Multi-Robot Adaptive Exploration and Mapping for Environmental Sensing Applications

Kian Hsiang Low

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

in the

Carnegie Institute of Technology

of the

Carnegie Mellon University

Thesis Committee

Pradeep K. Khosla
John M. Dolan
Jeff Schneider, RI, CMU
Alberto Elfes, JPL, NASA

Summer, 2009

Abstract

Recent research in robot exploration and mapping has focused on sampling hotspot fields, which often arise in environmental and ecological sensing applications. Such a hotspot field is characterized by *continuous, positively skewed, spatially correlated* measurements with the hotspots exhibiting extreme measurements and much higher spatial variability than the rest of the field.

To map a hotspot field of the above characterization, we assume that it is realized from non-parametric probabilistic models such as the Gaussian and log-Gaussian processes (respectively, GP and ℓGP), which can provide formal measures of map uncertainty. To learn a hotspot field map, the exploration strategy of a robot team then has to plan resource-constrained observation paths that minimize the uncertainty of a spatial model of the hotspot field. This exploration problem is formalized in a sequential decision-theoretic planning under uncertainty framework called the *multi-robot adaptive sampling problem* (MASP). So, MASP can be viewed as a sequential, non-myopic version of active learning. In contrast to finite-state Markov decision problems, MASP adopts a more complex but realistic continuous-state, non-Markovian problem structure so that its induced exploration policy can be informed by the complete history of continuous, spatially correlated observations for selecting paths. It is unique in unifying formulations of non-myopic exploration problems along the entire adaptivity spectrum, thus subsuming existing non-adaptive formulations and allowing the performance advantage of a more adaptive policy to be theoretically realized. Through MASP, it is demonstrated that a more adaptive strategy can exploit clustering phenomena in a hotspot field to produce lower expected map uncertainty. By measuring map uncertainty using the mean-squared error criterion, a MASP-based exploration strategy consequently plans adaptive observation paths that minimize the expected posterior map error or equivalently, maximize the expected map error reduction.

The time complexity of solving MASP (approximately) depends on the map resolution, which limits its practical use in large-scale, high-resolution exploration and mapping. This computational difficulty is alleviated through an information-theoretic approach to MASP (iMASP), which measures map uncertainty based on the entropy criterion instead. As a result, an iMASP-based exploration strategy plans adaptive observation paths that minimize the expected posterior map entropy or equivalently, maximize the expected entropy of observation paths. Unlike MASP, reformulating the cost-minimizing iMASP as a reward-maximizing dual problem causes its time complexity of being solved approximately to be independent of the map resolution and less sensitive to larger robot team size as demonstrated both analytically and empirically. Furthermore, this reward-maximizing dual transforms the widely-used non-adaptive maximum entropy sampling problem into a novel adaptive variant, thus improving the performance of the induced exploration policy.

One advantage stemming from the reward-maximizing dual formulations of MASP and iMASP is that they allow observation selection properties of the induced exploration policies to be realized for sampling the hotspot field. These properties include adaptivity, hotspot sampling, and wide-area coverage. We show that existing GP-based exploration strategies may not explore and map the hotspot field well with the selected observations because they are non-adaptive and perform only wide-area coverage. In contrast, the ℓGP-based exploration policies can learn a high-quality hotspot field map because they are adaptive and perform both wide-area coverage and hotspot sampling.

The other advantage is that even though MASP and iMASP are non-trivial to solve due to their continuous state components, the convexity of their reward-maximizing duals can be exploited to derive, in a computationally tractable manner, discrete-state monotone-bounding approximations and subsequently, approximately optimal exploration policies with theoretical performance guarantees. Anytime algorithms based on approximate MASP and iMASP are then proposed to alleviate the computational difficulty that arises from their non-Markovian structure.

It is of practical interest to be able to quantitatively characterize the "hotspotness" of an environmental field. We propose a novel "hotspotness" index, which is defined in terms of the spatial correlation properties of the hotspot field. As a result, this index can be related to the intensity, size, and diffuseness of the hotspots in the field.

We also investigate how the spatial correlation properties of the hotspot field affect the performance advantage of adaptivity. In particular, we derive sufficient and necessary conditions of the spatial correlation properties for adaptive exploration to yield no performance advantage.

Lastly, we develop computationally efficient approximately optimal exploration strategies for sampling the GP by assuming the Markov property in iMASP planning. We provide theoretical guarantees on the performance of the Markov-based policies, which improve with decreasing spatial correlation. We evaluate empirically the effects of varying spatial correlations on the mapping performance of the Markov-based policies as well as whether these Markov-based path planners are time-efficient for the transect sampling task.

Through the abovementioned work, this thesis establishes the following two claims: (1) adaptive, non-myopic exploration strategies can exploit clustering phenomena to plan observation paths that produce lower map uncertainty than non-adaptive, greedy methods; and (2) Markov-based exploration strategies can exploit small spatial correlation to plan observation paths which achieve map uncertainty comparable to that of non-Markovian policies using significantly less planning time.

Acknowledgments

I am eternally grateful to God for his love, provision, grace, peace, and guidance.

I would like to express my gratitude to my advisors, John M. Dolan and Pradeep K. Khosla, for providing timely advice and guidance during my Ph.D. studies. I would also like to thank the other committee members, Jeff Schneider and Alberto Elfes, for their valuable feedback.

My stay in Pittsburgh would never be complete without my wife, Cocoa Yeo, who has showered me with overwhelming love, care, and concern.

For my fellow brothers and sisters in Abundant Life cell group and Oakland International Fellowship, I want to thank all of you[1] for making us feel at home. I'll miss the fellowship time that we have together during bible studies, trail hikes, cabin stays, and potluck sessions with savory Singaporean food.

Lastly, I will always be grateful to my family for their constant support when I needed them most.

[1]In particular, Rodney and Suzanne, Samuel and Hwee Leng, Shawn and Kelly, Daniel and Christine, Philip and Pollyn, Paul and Linda Bucknell, Wei Siong, Brian Lim, Woon Chiat, Brian and Sharon Conner.

Contents

List of Figures **xi**

List of Tables **xv**

1 Introduction **1**

1.1 Motivation 1

1.2 Objective 3

1.3 Contributions 4

1.3.1 Formalization of multi-robot adaptive sampling problem 4

1.3.2 Formalization of information-theoretic multi-robot adaptive sampling problem 6

1.3.3 Exploration strategies for learning hotspot field maps 7

1.3.4 Approximately optimal exploration strategies with performance guarantees 9

1.3.5 Quantifying "hotspotness" 10

1.3.6 Effects of spatial correlation on performance advantage of adaptivity 10

1.3.7 Exploiting small spatial correlation with fast Markov-based exploration strategies 11

1.3.8 Summary of contributions 12

2 Related Work on Exploration Strategies **17**

2.1 Wide-area coverage vs. feature sampling 17

2.2 Design-based vs. model-based strategies 18

2.3 Adaptive vs. non-adaptive sampling strategies 19

2.4 Greedy vs. non-myopic path planning strategies 20

2.5 Single- vs. multi-robot strategies 21

3 Multi-Robot Adaptive Sampling Problem (MASP) **23**

3.1 Terminology and Notation 24

3.2 Problem Formulations 26

3.2.1 Objective function 26

3.2.2 Value function 27

3.2.3 Adaptive exploration 28

3.2.4 Non-adaptive exploration 30

3.3 Performance Advantage of Adaptive Exploration 32

3.4 Dual Formulations 34

3.5 Learning the Hotspot Field Map 37

3.5.1 Gaussian process (GP) 37

3.5.2 Log-Gaussian process (ℓGP) 42

4 Information-Theoretic Multi-Robot Adaptive Sampling Problem (iMASP) **47**

4.1 Problem Formulations 48

4.1.1 Objective function 48

4.1.2 Value function 49

4.1.3 Adaptive and non-adaptive exploration 49

4.2 Dual Formulations 50

4.3 Learning the Hotspot Field Map 53

4.3.1 Gaussian process (GP) 53
4.3.2 Log-Gaussian process (ℓGP) 55

5 Value-Function Approximations 57
5.1 Strictly Adaptive Exploration 57
5.2 Related Work on Sequential Decision-Theoretic Planning with Continuous States 60
5.3 Approximately Optimal Exploration 63
5.4 Bounds on Performance Advantage of Adaptive Exploration 68
5.5 Real-Time Dynamic Programming 69
5.5.1 Preprocessing of heuristic bounds 70
5.5.2 Anytime algorithms 73

6 Experiments and Discussion 77
6.1 Performance Metrics 79
6.2 Test Results 80
6.2.1 Plankton density data 80
6.2.2 Potassium distribution data 81
6.2.3 Summary of test results 82

7 Quantifying "Hotspotness" 83
7.1 Index of "Hotspotness" 84
7.2 Effects of Spatial Correlation on Hotspot Characteristics and "Hotspotness" 88
7.3 Application: Phosphorus Distribution Field 88
7.4 Generalized Index of "Hotspotness" 89

8 Effects of Spatial Correlation on Performance Advantage of Adaptivity 93
8.1 Multi-Stage MASP(1) and iMASP(1) 94

8.2 2-Stage iMASP(1) 94
8.2.1 Exact closed-form solution for adaptive iMASP(1) 95
8.2.2 Exact closed-form solution for non-adaptive iMASP(n) 96
8.2.3 Performance advantage of adaptive exploration 97

9 Fast Information-Theoretic Path Planning with Deterministic MDP for Active Sampling of Gaussian Process 105

9.1 Deterministic Non-Markovian iMASP(1) 106
9.2 Transect Sampling Task 107
9.3 Deterministic Markov Decision Process (DMDP) 109
9.4 Deterministic Markov Decision Process with Factored Reward (DMDP+FR) 115
9.5 Experiments and Discussion 120
9.5.1 Performance metrics 121
9.5.2 Test results 121
9.5.3 Summary of test results 134

10 Conclusion and Future Work 137

10.1 Summary of Contributions 137
10.2 Future Work 139

Bibliography 141

A Proofs 151

A.1 Theorem 3.3.1 151
A.2 Theorem 3.4.1 153
A.3 Lemma 3.5.1 154
A.4 Lemma 3.5.2 155
A.5 Theorem 4.2.1 157

A.6 Equation 4.8 158
A.7 Equation 4.10 159
A.8 Generalized Jensen and Edmundson-Madansky bounds 160
A.9 Lemma 5.3.1 163
A.10 Theorem 5.3.1 165
A.11 Theorem 5.3.2 166
A.12 Monotonicity of Lower Heuristic Bound 168
A.13 Monotonicity of Upper Heuristic Bound 168
A.14 Theorem 8.1.1 169
A.14.1 MASP 170
A.14.2 *i*MASP 173
A.15 Equation 8.2 173
A.16 Lemma 9.3.1 174
A.17 Theorem 9.3.2 177
A.18 Theorem 9.3.3 178
A.19 Lemma A.18.1 179

List of Figures

1.1 Plankton density (chlorophyll-a) field of Chesapeake Bay showing hotspots to the left. 2

3.1 Illustration of strictly adaptive ($k = 1$), partially adaptive ($k = 2$, $k = 5$), and non-adaptive ($k = 10$) exploration strategies: the number of new locations n to be explored is 10. Each box denotes a stage consisting of the k locations aligned above it. For example, the strictly adaptive strategy spans 10 stages and selects 1 new location per stage. On the other hand, the non-adaptive strategy selects all 10 new locations in a single stage. Please refer to Definition 3.1.2 for a formal characterization of adaptive and non-adaptive exploration strategies. 25

3.2 1D sample frequency distributions/histograms of the plankton density field (Fig. 1.1) measurements (a) before taking log, and (b) after taking log. . . . 40

3.3 Hotspot field simulation via the (a) ℓGP, and (b) GP with their respective 1D sample frequency distributions/histograms. 43

6.1 (a) chl-a field with prediction error maps for (b) strictly adaptive $\underline{\pi}^{\frac{1}{k}}$ and (c) non-adaptive π^n: 20 units (white circles) are randomly selected as prior data. The robots start at locations marked by '×'s. The black and gray robot paths are produced by $\underline{\pi}^{\frac{1}{k}}$ and π^n respectively. (d-f) K field with prediction error maps for $\underline{\pi}^{\frac{1}{k}}$ and π^n. 78

7.1 Graphical interpretation of the degree of "hotspotness" H_θ measuring the bounded area covered with '+'s. Refer to definition 7.1.1 for more details. . . 86

7.2 Hotspot field simulations via ℓGP with varying hyperparameters producing different characteristics of hotspots and degrees of "hotspotness" H_0. See text for explanation. 87

7.3 Phosphorus distribution field with $H_0 = 1.7142$, $H_0 = 1.2486$, and $H_0 = 0.9542$ for the sub-regions boxed with a dashed red line, a dotted green line, and a solid blue line respectively. 89

7.4 Graphical interpretations of H_θ^1 and H_θ^2 measuring the bounded areas covered, respectively, by '+'s and '×'s. See the proof of Theorem 7.4.1 for more details. 92

8.1 Graph of performance advantage $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0)$ vs. length-scale ℓ for varying signal variances σ_s^2. 102

9.1 Transect sampling task over a 25 m × 150 m temperature field discretized into a 5 × 30 grid of sampling locations (white dots). 108

9.2 Temperature field discretized into a 5 × 30 grid of sampling locations (white dots). 121

9.3 Temperature fields with varying horizontal and vertical length-scales. 122

9.4 Performance comparison of DMDP-based, DMDP+FR-based, and non-Markovian greedy policies for the case of 1 robot. 126

9.5 Squared relative error maps for field 3 with 1 robot showing white grid cells sampled by the (a) DMDP-based policy $\widetilde{\pi}$ and (b) non-Markovian greedy policy π_G. 127
9.6 Performance comparison of DMDP-based, DMDP+FR-based, and non-Markovian greedy policies for the case of 2 robots. 128
9.7 Squared relative error maps for field 2 with 2 robots showing white grid cells sampled by the (a) DMDP+FR-based policy $\widehat{\pi}$, and (b) DMDP-based policy $\widetilde{\pi}$ and non-Markovian greedy policy π_G. 129
9.8 Performance comparison of DMDP-based, DMDP+FR-based, and non-Markovian greedy policies for the case of 3 robots. 131
9.9 Squared relative error maps for field 3 with 3 robots showing white grid cells sampled by the (a) DMDP-based policy $\widetilde{\pi}$, and (b) non-Markovian greedy policy π_G. 132
9.10 Graph of time taken to derive policy vs. number of robots k for temperature field 4. 135

List of Tables

2.1 Qualitative comparison of directed exploration strategies (WC: Wide-area Coverage, FS: Feature Sampling, DB: Design-Based, MB: Model-Based, SA: Strictly Adaptive, PA: Partially Adaptive, NA: Non-Adaptive, GR: Greedy, NM: Non-Myopic, NO: Non-Optimized, SR: Single-Robot, MR: Multi-Robot). 18

5.1 Comparison of structural constraints on time-dependent MDPs with respect to the continuous state. 61

6.1 Performance comparison of information-theoretic policies for chl-a and K fields: 1R (2R) denotes 1 (2) robots. 81

Chapter 1

Introduction

1.1 Motivation

The problem of exploring and mapping an unknown environmental field is a central issue in mobile robotics. Typically, it requires sampling the entire terrain (Burgard *et al.*, 2005; Choset, 2001). With limited (e.g., point-based) robot sensing range, a complete coverage becomes impractical in terms of resource costs (e.g., energy consumption) if the environmental field to be explored is large with only a few small-scale features of interest, or *hotspots*. Such a *hotspot field* (see, for example, Fig. 1.1) characterizes two real-world application domains:

- planetary exploration such as antarctic meteorite search (Apostolopoulos *et al.*, 2000), geologic site survey (Castano *et al.*, 2003; Glass and Briggs, 2003; Estlin *et al.*, 1999, 2005), and prospecting for mineral deposits (Low *et al.*, 2007) or localized methane sources on Mars (Formisano *et al.*, 2004; Krasnopolsky *et al.*, 2004), and

- environmental and ecological sensing such as precision agriculture (Pedersen *et al.*, 2006), monitoring of ocean phenomena (plankton bloom, upwelling) (Fiorelli *et al.*, 2006; Leonard *et al.*, 2007), forest ecosystems (Batalin *et al.*, 2004; Rahimi *et al.*,

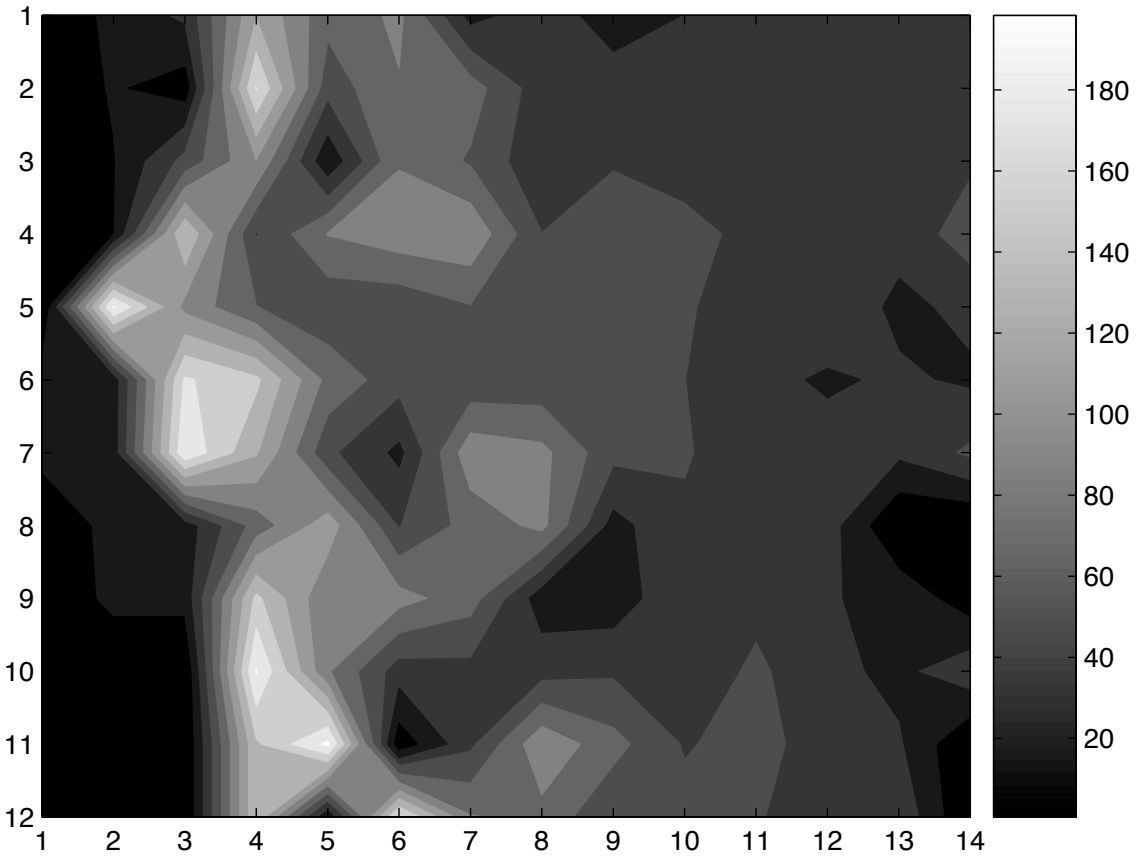


Figure 1.1: Plankton density (chlorophyll-a) field of Chesapeake Bay showing hotspots to the left.

2004), rare species (Thompson, 2004), pollution (Chang *et al.*, 2007; Long and Wilson, 1997), or contamination (Englund and Heravi, 1994).

In particular, the hotspot field is characterized by *continuous, positively skewed, spatially correlated* measurements with the hotspots exhibiting extreme measurements and much higher spatial variability than the rest of the field. So, to accurately map the field, the hotspots have to be sampled at a higher resolution.

The hotspot field discourages static sensor placement (Guestrin *et al.*, 2005) because a large number of sensors has to be positioned to detect and refine the sampling of hotspots. If these static sensors are not placed in any hotspot initially, they cannot reposition by themselves to locate one. Furthermore, these static sensors are not capable of sample return or withdrawal from harsh environmental conditions, and they possess limited sensory, computational, and communication capabilities. In contrast, a small robot team is capable of performing high-resolution sampling of the hotspots due to its mobility. Hence, it is desirable to build a mobile robot team that can actively explore to map a hotspot field.

1.2 Objective

An important issue in designing such a robot team is the *exploration strategy* for learning a hotspot field map. This thesis aims to address this issue:

> How does a robot team plan resource-constrained observation paths to minimize the map uncertainty of a hotspot field?

The related work pertaining to this objective will be discussed in Chapter 2. To achieve this objective, we believe that such an exploration strategy should be designed to

- exploit the environmental structure: the large hotspot field is characterized by *continuous, positively skewed, spatially correlated* measurements, and contains a few *small-scale* hotspots with *extreme, highly-varying* measurements;

- maximize science return: the environmental scientist deploying the robot team expects the scientific exploration to yield a *high-quality* map that, in particular, identifies and represents the localized *features of interest* (i.e., hotspots) well.

These two factors are considered in the development of our proposed exploration strategies as described next.

1.3 Contributions

The work in this thesis makes the following two claims:

> Adaptive, non-myopic exploration strategies can exploit clustering phenomena to plan observation paths that produce lower map uncertainty than non-adaptive, greedy methods.

> Markov-based exploration strategies can exploit small spatial correlation to plan observation paths which achieve map uncertainty comparable to that of non-Markovian policies using significantly less planning time.

Both claims are substantiated by the following novel contributions listed in Sections 1.3.1 to 1.3.6 and Section 1.3.7, respectively.

1.3.1 Formalization of multi-robot adaptive sampling problem

We cast the exploration task as a sequential (i.e., stagewise) decision-theoretic planning problem and call it the *multi-robot adaptive sampling problem* (MASP) (Section 3.2). That is, the robot team's goal is to minimize the map uncertainty over a given finite planning horizon (e.g., task duration or observation path length). Hence, MASP can be viewed as a generalization of active learning (Cohn *et al.*, 1996; Roy and McCallum, 2001) due to its

sequential nature. Very often, a sequential decision-theoretic planning problem is framed as a finite-state *Markov decision process* (MDP). To learn an accurate map of the hotspot field, the MDP-based exploration problem has to exploit its environmental structure (Section 1.2). However, it cannot do this because its discrete-state, Markov assumptions are violated by the continuous, spatially correlated field measurements. To fully exploit the environmental structure, MASP adopts a more complex but realistic *continuous-state*[1], *non-Markovian* problem structure so that (a) its induced exploration policy can be informed by the complete history of continuous, spatially correlated observations for selecting observation paths, and (b) theoretical bounds can be established for the expected map uncertainty achieved by solving MASP, as explained further in Section 1.3.4.

In particular, we endow the induced exploration policy from solving MASP with *non-myopic* and *adaptive* observation selection properties to improve its performance. Intuitively, the choice of next immediate observation(s) under a non-myopic policy looks beyond the next immediate observation selection(s) while the choice of new observation(s) under an adaptive policy depends on past observation selections. By looking ahead, a non-myopic policy can potentially perform better than a greedy one as explained in Section 2.4. On the other hand, a more adaptive policy can be obtained not by depending on a longer history of observations[2] but by increasing its frequency of observation selection. Given a fixed observation path length for each robot, this then can be achieved by refining the sequential (i.e., stagewise) observation selection to fewer number of new observations per stage but over a longer planning horizon (i.e., greater number of stages) (Section 3.1). If the frequency of observation selection is reduced to a single stage instead, a non-adaptive policy results. In this case, all observations have to be chosen in a single stage and therefore cannot depend

[1]In this thesis, each observation path is assumed to be a finite collection of *continuous-valued* field measurements sampled at *discrete* locations. In practice, this assumption is reasonable because the sampling rate of a sensor is usually finite. Without loss of generality, we assume that the locations to be selected by the observation paths correspond to that of a discretized grid (Section 3.1).

[2]This implies it is "more" non-Markovian instead.

on past observation selections. As shown later in Section 3.2, the seemingly contrasting non-myopic and adaptive properties cannot be treated exclusively or decoupled in an exploration problem because the degree of adaptivity in a policy dictates how a non-myopic exploration problem is to be formulated.

The formalization of MASP is unique in its unification of formulations of non-myopic exploration problems with varying adaptivity. As a result, this unifying framework covers the entire adaptivity spectrum, thus subsuming various existing non-adaptive problem formulations (Chapter 2). It is useful in allowing the performance of induced exploration policies of varying adaptivity to be theoretically analyzed; the performance advantage of a more adaptive exploration policy can thus be realized (Section 3.3). Through MASP, it is demonstrated that a more adaptive policy can exploit clustering phenomena in a hotspot field to produce lower expected map uncertainty.

For MASP, the map uncertainty is measured using the mean-squared error criterion. Consequently, solving MASP involves planning observation paths that produce the least expected posterior map error/uncertainty; the original MASP is therefore a cost-minimizing problem. We then provide a reward-maximizing dual formulation of MASP that, when solved, achieves the largest possible expected reduction in map error/uncertainty (Section 3.4). We show that the exploration objectives of these two problem formulations are equivalent. That is, the induced optimal exploration policies from solving the cost-minimizing and reward-maximizing MASPs coincide.

1.3.2 Formalization of information-theoretic multi-robot adaptive sampling problem

The MASP is beset by a serious computational drawback due to its measure of map uncertainty using the mean-squared error criterion. Consequently, the time complexity of solving

MASP approximately[3] depends on the map resolution, which limits the practical use of MASP-based approximation algorithms in large-scale, high-resolution exploration and mapping (Chapter 5).

This computational difficulty is alleviated through an information-theoretic approach to MASP (iMASP) (Section 4.1), which measures map uncertainty based on the entropy criterion instead. Unlike MASP, reformulating the cost-minimizing iMASP as a reward-maximizing problem (Section 4.2) causes its time complexity of being solved approximately[4] to be independent of the map resolution and less sensitive to larger robot team size as demonstrated both analytically (Section 5.5.2) and empirically (Chapter 6). We also show the equivalence between the cost-minimizing and reward-maximizing iMASPs (Section 4.2). That is, their induced optimal exploration policies coincide. Beyond its computational gain, iMASP retains the beneficial properties of MASP.

Additional contributions stemming from this reward-maximizing formulation include (a) transforming the commonly-used non-adaptive maximum entropy sampling problem (Shewry and Wynn, 1987) into a novel adaptive variant, thus improving the performance of the induced exploration policy (Section 4.2); and (b) given an assumed environment model (e.g., occupancy grid map), establishing sufficient conditions that, when met, guarantee adaptivity provides no benefit (Section 4.3.1).

1.3.3 Exploration strategies for learning hotspot field maps

Non-parametric probabilistic models such as *Gaussian* and *log-Gaussian processes* (respectively, GP and ℓGP) are commonly used to map environmental fields (Cressie, 1993). These models offer the advantage of providing formal measures of map uncertainty. By reformulating the cost-minimizing MASP and iMASP as reward-maximizing problems, observation se-

[3]See Section 1.3.4 to understand why it is computationally intractable to solve MASP exactly.
[4]See Section 1.3.4 to understand why it is computationally intractable to solve iMASP exactly.

lection properties of the induced exploration policies can be realized when they are applied to sampling GP and ℓGP. These properties include adaptivity, wide-area coverage, and hotspot sampling. In particular, wide-area coverage and hotspot sampling assess highly uncertain map regions differently: the former considers sparsely sampled areas to be of high uncertainty while the latter expects areas of high uncertainty to contain extreme, highly-varying measurements. These two, at times contrasting, properties trigger the exploration-exploitation dilemma where wide-area coverage parallels exploration of uncharted terrain to locate previously undetected hotspots and hotspot sampling coincides with exploitation to maximize the sampling at hotspots. Hence, they are both necessary for learning a high-quality hotspot field map.

Existing exploration strategies (Alvarez *et al.*, 2007; Meliou *et al.*, 2007; Singh *et al.*, 2007) have used GP to model environmental fields and are devised to select observations for reducing the uncertainty of the GP model. Using the reward-maximizing MASP and iMASP, we can show that they are, however, non-adaptive and perform only wide-area coverage (Sections 3.5.1 and 4.3.1). So, a hotspot field may not be reconstructed well with the selected observations from wide-area coverage because the under-sampled hotspots with extreme, highly-varying measurements can contribute considerably to the map uncertainty. If ℓGP is used to model the hotspot field instead, it can be shown using the reward-maximizing MASP and iMASP that the induced exploration policies are adaptive, and perform both wide-area coverage and hotspot sampling (Sections 3.5.2 and 4.3.2). As demonstrated empirically (Chapter 6), the adaptive exploration policies for sampling ℓGP select observation paths through hotspots to produce lower map uncertainty.

1.3.4 Approximately optimal exploration strategies with performance guarantees

Sequential decision-theoretic planning problems such as MASP and iMASP are generally non-trivial to solve. Although MASP and iMASP can be solved exactly via dynamic programming, this is computationally intractable due to their (a) continuous state components and (b) non-Markovian structure (Section 1.3.1). As elaborated below, we exploit the problem structure for sampling ℓGP to derive approximately optimal exploration policies in a computationally tractable manner.

Recent solution techniques of continuous-state, reward-maximizing MDPs (Boyan and Littman, 2001; Feng *et al.*, 2004; Li and Littman, 2005; Marecki *et al.*, 2007; Kveton and Hauskrecht, 2006; Kveton *et al.*, 2006) constrain the transition, reward, and optimal value functions to specific function families (e.g., discrete, piecewise-constant and linear, etc). In contrast, we assume the reward function merely to be convex, and do not restrict the form of transition function. Consequently, the optimal value functions are convex.

To handle continuous states, we show that the reward-maximizing MASP and iMASP are convex, which allows discrete-state monotone-bounding approximations to be developed (Section 5.3). Consequently, we can provide theoretical guarantees on the performance of approximately optimal vs. optimal adaptive policies (Section 5.3), and establish theoretical bounds quantifying the performance advantage of optimal adaptive over non-adaptive policies (Section 5.4).

Although it is computationally tractable to solve the approximate MASP and iMASP exactly, their non-Markovian structure causes the state size to grow exponentially with the number of stages. To alleviate this computational difficulty, anytime algorithms are proposed based on the approximate MASP and iMASP, which can guarantee their policy performance in real time (Section 5.5). As demonstrated analytically, the time complexity of the iMASP-based anytime algorithm is independent of map resolution and less sensitive to increasing

robot team size as compared to the MASP-based algorithm.

1.3.5 Quantifying "hotspotness"

It is of practical interest to be able to quantitatively characterize the "hotspotness" of an environmental field. In this manner, environmental fields of varying degrees of "hotspotness" can be prescribed accordingly, that is, by assigning high degrees of "hotspotness" to fields with pronounced hotspots and low degrees of "hotspotness" to smoothly-varying fields.

We propose a novel "hotspotness" measure, which is defined in terms of the spatial correlation properties of the hotspot field (Section 7.1). Specifically, by assuming the hotspot field to vary as a realization of the ℓGP (Sections 3.5.2 and 4.3.2), its spatial correlation properties can be represented by the hyperparameters of the ℓGP covariance structure. This then allows the proposed "hotspotness" index to be defined using the hyperparameters. Through the use of the hyperparameters, we discuss how the "hotspotness" index can be related to the intensity, size, and diffuseness of the hotspots in the environmental field (Section 7.2). The "hotspotness" index is applied to a real-world phosphorus distribution field (Section 7.3).

1.3.6 Effects of spatial correlation on performance advantage of adaptivity

We investigate how the spatial correlation properties of the hotspot field (in particular, the length-scale hyperparameter of the ℓGP covariance structure) affect the performance advantage of adaptive exploration. We first show that for white-noise process fields or constant fields, multi-stage adaptive MASP and iMASP provide no performance advantage under the mean-squared error or entropy criterion, respectively (Section 8.1). Then, we show that the performance advantage of the 2-stage adaptive iMASP is zero if and only if

the field is constant or a white-noise process (Section 8.2). Note that the contrapositive of this second result implies the performance advantage is positive if and only if the length-scale is non-extreme. Lastly, we illustrate that the performance advantage of the 2-stage adaptive *i*MASP improves with decreasing noise-to-signal variance ratio and peaks at some intermediate length-scale.

1.3.7 Exploiting small spatial correlation with fast Markov-based exploration strategies

The *i*MASP for sampling GP can be reduced to a non-Markovian, deterministic planning problem (Section 9.1). Due to its non-Markovian structure, the state size grows exponentially with the number of stages. Furthermore, the time complexity of evaluating each entropy-based stagewise reward in *i*MASP depends cubically on the length of the history of observations, which limits the practical use of its approximation algorithm in *in situ* real-time, high-resolution active sampling. This latter computational difficulty also plagues the widely-used non-Markovian greedy algorithm as it is a single-staged variant of *i*MASP.

We develop computationally efficient exploration strategies for sampling the GP by assuming the Markov property in *i*MASP planning. The resulting information-theoretic path planning problem can be cast as a *deterministic Markov decision process* (DMDP) (Section 9.3). We analyze the time complexity of solving the DMDP-based path planning problem, and show analytically that it scales better than the non-Markovian greedy algorithm with increasing number of planning stages. We also provide a theoretical guarantee on the performance of the DMDP-based policy for the case of a single robot, which, in particular, improves with decreasing spatial correlation.

Unfortunately, the performance guarantee of the DMDP-based policy cannot be generalized to the case of multiple robots unless we impose more restrictive assumptions on the GP covariance structure. However, we can obtain a similar form of performance guarantee by

factoring the stagewise reward (Section 9.4), which essentially imposes a conditional independence assumption. The resulting path planning problem is therefore framed as a DMDP with factored reward (DMDP+FR). In terms of time complexity, we show analytically that it scales better than the DMDP-based algorithm with increasing number of robots.

Through the transect sampling task (Section 9.2), we investigate empirically the effects of varying spatial correlations on the mapping performance of the Markov-based policies as well as whether these Markov-based path planners are time-efficient for *in situ* real-time, high-resolution active sampling (Section 9.5).

1.3.8 Summary of contributions

To summarize, the work in this thesis provides the following novel contributions and is organized as follows:

1. Formalization of MASP (Chapter 3). MASP formalizes the exploration problem in a sequential decision-theoretic planning under uncertainty framework. It is unique in unifying formulations of non-myopic exploration problems along the entire adaptivity spectrum, thus allowing it to subsume various existing non-adaptive problem formulations. Consequently, it allows the performance of induced exploration policies of varying adaptivity to be theoretically analyzed and the performance advantage of a more adaptive policy to be realized. Through MASP, it is demonstrated that a more adaptive policy can exploit clustering phenomena in a hotspot field to produce lower expected map uncertainty. MASP measures the map uncertainty using the mean-squared error criterion. Consequently, solving MASP involves planning observation paths that produce the least expected posterior map error, and is therefore cost-minimizing. We provide a reward-maximizing dual formulation of MASP that, when solved, achieves the largest possible expected reduction in map error. We show that the exploration objectives of these two problem formulations are equivalent.

2. Formalization of iMASP (Chapter 4). The time complexity of solving MASP (approximately) depends on the map resolution due to its measure of map uncertainty using the mean-squared error criterion, which limits its practical use in large-scale, high-resolution exploration and mapping. This computational difficulty is alleviated through iMASP, which measures map uncertainty based on the entropy criterion instead. Unlike MASP, reformulating the cost-minimizing iMASP as a reward-maximizing problem causes its time complexity of being solved approximately to be independent of the map resolution and less sensitive to larger robot team size as demonstrated both analytically (Section 5.5.2) and empirically (Chapter 6). Furthermore, this reward-maximizing dual transforms the widely-used non-adaptive maximum entropy sampling problem into a novel adaptive variant, thus improving the performance of the induced exploration policy.

3. Exploration strategies for learning hotspot field maps (Chapters 3 and 4). The reward-maximizing MASP and iMASP allow observation selection properties of the induced exploration policies to be realized for sampling GP and ℓGP. For example, we can show that existing exploration strategies utilizing GP may not reconstruct the hotspot field well with the selected observations because they are non-adaptive and perform only wide-area coverage (Sections 3.5.1 and 4.3.1, and Chapter 6). By modeling with ℓGP, the induced exploration policies can learn a high-quality hotspot field map because they are adaptive and perform both wide-area coverage and hotspot sampling (Sections 3.5.2 and 4.3.2, and Chapter 6).

4. Approximately optimal exploration strategies with performance guarantees (Chapter 5). We exploit the problem structure for sampling ℓGP to derive approximately optimal exploration policies in a computationally tractable manner. To handle continuous states, the convexity of reward-maximizing MASP and iMASP allows discrete-state monotone-bounding approximations to be developed. Consequently, we can provide

theoretical guarantees on the performance of the approximately optimal vs. the optimal adaptive policies, and establish theoretical bounds quantifying the performance advantage of optimal adaptive over non-adaptive policies. We then propose anytime algorithms based on approximate MASP and iMASP to alleviate the computational difficulty that arises from their non-Markovian structure.

5. Quantifying "hotspotness" (Chapter 7). We propose a "hotspotness" measure, which is defined in terms of the spatial correlation properties of the hotspot field. We discuss how the "hotspotness" index can be related to the intensity, size, and diffuseness of the hotspots in the field.

6. Effects of spatial correlation on performance advantage of adaptivity (Chapter 8). We investigate how the spatial correlation properties of the hotspot field affect the performance advantage of adaptive exploration, and obtain the following three results: (a) for white-noise process fields or constant fields, multi-stage adaptive MASP and iMASP provide no performance advantage under the mean-squared error and entropy criteria, respectively; (b) the 2-stage adaptive iMASP yields no performance advantage if and only if the field is constant or a white-noise process; (c) the performance advantage of the 2-stage adaptive iMASP improves with decreasing noise-to-signal variance ratio and peaks at some intermediate length-scale.

7. Exploiting small spatial correlation with fast Markov-based exploration strategies (Chapter 9). We develop computationally efficient exploration strategies for sampling the GP by assuming the Markov property in iMASP planning. The resulting path planning problem can be cast as a DMDP. In terms of time complexity, it scales better than the non-Markovian greedy algorithm with increasing number of planning stages. We provide a theoretical guarantee on the performance of the DMDP-based policy for the single-robot case, which improves with decreasing spatial correlation. To generalize the

performance guarantee to the multi-robot case, we have to factor the entropy-based stagewise reward. In terms of time complexity, the DMDP+FR-based algorithm scales better than the DMDP-based one with increasing number of robots.

Chapter 2

Related Work on Exploration Strategies

Our proposed exploration strategies (Low *et al.*, 2008, 2009) are the first in the class of model-based strategies to perform both wide-area coverage and hotspot sampling, and cover the entire adaptivity spectrum. In contrast, all other model-based strategies are non-adaptive and achieve only wide-area coverage (Table 2.1). Our strategies can also plan non-myopic multi-robot paths, which are more desirable than greedy or single-robot paths. These characteristics distinguish our approach from the existing robot exploration strategies and are discussed in greater detail with the related work below.

2.1 Wide-area coverage vs. feature sampling

In contrast to random exploration of the environmental field (McCartney and Sun, 2000), directed exploration selects robot paths to observe regions of high uncertainty. One such class of strategies emphasizes *wide-area coverage* (Alvarez *et al.*, 2007; Leonard *et al.*, 2007; Meliou *et al.*, 2007; Popa *et al.*, 2006; Popa and Lewis, 2008; Singh *et al.*, 2007; Zhang and Sukhatme, 2007), which considers sparsely sampled (i.e., largely unexplored) areas to be of high uncertainty. On the other hand, directed exploration strategies (Batalin *et al.*, 2004; Low *et al.*, 2007; Rahimi *et al.*, 2004; Singh *et al.*, 2006) that focus on *feature sampling*

Table 2.1: Qualitative comparison of directed exploration strategies (WC: Wide-area Coverage, FS: Feature Sampling, DB: Design-Based, MB: Model-Based, SA: Strictly Adaptive, PA: Partially Adaptive, NA: Non-Adaptive, GR: Greedy, NM: Non-Myopic, NO: Non-Optimized, SR: Single-Robot, MR: Multi-Robot).

Characteristic / Exploration Strategy	WC	FS	DB	MB	SA	PA	NA	GR	NM	NO	SR	MR
(Low *et al.*, 2007)		×	×			×				×		×
(Singh *et al.*, 2006)		×	×			×				×		×
(Batalin *et al.*, 2004; Rahimi *et al.*, 2004)		×	×			×				×	×	
(Leonard *et al.*, 2007)	×			×			×			×		×
(Meliou *et al.*, 2007)	×			×			×	×			×	
(Popa *et al.*, 2006; Popa and Lewis, 2008)	×			×			×	×			×	
(Singh *et al.*, 2007)	×			×			×	×				×
(Alvarez *et al.*, 2007)	×			×			×		×			×
(Zhang and Sukhatme, 2007)	×			×			×		×		×	
MASP (Low *et al.*, 2008)	×	×		×	×	×	×		×			×
***i*MASP** (Low *et al.*, 2009)	×	×		×	×	×	×		×			×

(e.g., hotspots) expect areas of high uncertainty to contain highly-varying measurements. As a result, they tend to produce clustered observations, while the former strategies tend to spread the observations evenly across the environmental field. In contrast, we propose a formal, principled approach that can tackle both tasks simultaneously (i.e., by directing exploration towards sparsely sampled areas and hotspots).

2.2 Design-based vs. model-based strategies

In design-based strategies (Batalin *et al.*, 2004; Low *et al.*, 2007; McCartney and Sun, 2000; Rahimi *et al.*, 2004; Singh *et al.*, 2006), the selection of sampling locations for exploration is constrained by the sampling design, which is not devised to consider resource costs. As a result, the locations have to be chosen by the strategy first before minimizing the resource costs to sample them. This entails "sunk" costs in motion; the exploration paths have to traverse terrain that does not require sampling to reach the selected locations. Furthermore, some of these strategies (Batalin *et al.*, 2004; Rahimi *et al.*, 2004; Singh *et al.*, 2006) require multiple "passes" through the region of interest such that new locations are adaptively

selected and sampled in each pass. Note that the cost efficiency of these strategies improves when the cost of sampling/sensing each selected location significantly outweighs the cost of moving to this location (e.g., mineral prospecting task of Low *et al.* (2007)). Modifying the strategy to involve resource costs may invalidate the estimators associated with the strategy. For example, in simple random sampling, if the selection of sampling locations is subject to costs, the associated sample mean estimator becomes biased.

We instead use model-based strategies (Alvarez *et al.*, 2007; Leonard *et al.*, 2007; Meliou *et al.*, 2007; Popa *et al.*, 2006; Popa and Lewis, 2008; Singh *et al.*, 2007; Zhang and Sukhatme, 2007), which assume a certain model for the environmental field and select observations to reduce its uncertainty. Resource cost minimization or constraints may be applied to the selection process and the resulting exploration strategy is optimal subject to these constraints. In contrast to the strategies in (Leonard *et al.*, 2007; Popa *et al.*, 2006; Popa and Lewis, 2008) that use a parametric model, our approach utilizes a non-parametric model, which does not require any assumptions on the distribution underlying the observed sampling data. In particular, we model the environmental field as a stochastic spatial process (i.e., Gaussian process (Alvarez *et al.*, 2007; Meliou *et al.*, 2007; Singh *et al.*, 2007) and log-Gaussian process), which contrasts with the use of local models (e.g., locally weighted linear regression) to estimate the field in (Zhang and Sukhatme, 2007).

2.3 Adaptive vs. non-adaptive sampling strategies

Adaptive sampling refers to sampling strategies (Batalin *et al.*, 2004; Low *et al.*, 2007; Rahimi *et al.*, 2004; Singh *et al.*, 2006) in which the sequential policy for selecting new locations to be included in the robot paths depends on the past observations taken during exploration. On the other hand, non-adaptive sampling strategies (Alvarez *et al.*, 2007; Leonard *et al.*, 2007; McCartney and Sun, 2000; Meliou *et al.*, 2007; Popa *et al.*, 2006; Popa and Lewis, 2008; Singh

et al., 2007; Zhang and Sukhatme, 2007) have no such dependence and can therefore select the robot paths prior to exploration. When the environmental field is smoothly varying, non-adaptive sampling strategies are known to perform well (Singh *et al.*, 2006). However, if the environment contains hotspots, adaptive sampling can exploit the clustering phenomena to map the environmental field with lower uncertainty than non-adaptive sampling. In contrast to the abovementioned schemes, the adaptivity of our proposed strategies can be varied; a more adaptive strategy decreases expected map uncertainty of the environmental field.

2.4 Greedy vs. non-myopic path planning strategies

In contrast to greedy strategies (Meliou *et al.*, 2007; Popa *et al.*, 2006; Popa and Lewis, 2008; Singh *et al.*, 2007), our strategies generate non-myopic observation paths (Alvarez *et al.*, 2007; Zhang and Sukhatme, 2007). Existing myopic/greedy strategies for reducing map uncertainty pose the following problem for the exploration task: to minimize the map uncertainty, a greedy strategy only considers the next immediate observation(s) to make and ignores future/subsequent observation selections. Greedy observation paths can thus be obtained by executing this strategy repeatedly. But, the choice of the next immediate observation(s) may influence how much the map uncertainty can be reduced by future observation selections, which a greedy strategy fails to take into account. This is especially true for robot path planning because the previously selected observations in the paths constrain the possible choices of subsequent observations. As a result, greedy observation paths may incur higher map uncertainty than optimal non-myopic paths.

In general, non-myopic paths approximate the optimal trajectories better, but incur higher computational cost. To reduce computational expense, the non-myopic strategy of Alvarez *et al.* (2007) utilizes a heuristic search technique to derive approximately optimal paths without guarantees on the path quality. The non-myopic strategy of Zhang and Sukhatme

(2007) assumes that the uncertainty reduction from observing a new location is independent of other observations in the robot path. As a result, the uncertainty reduction due to an observation path is just the sum of uncertainty reductions from observing individual locations in the path and can be computed using breadth-first search. This assumption is violated by the presence of spatially correlated measurements in an environmental field, which severely limits its real-world use.

The design-based strategies (Batalin *et al.*, 2004; Low *et al.*, 2007; McCartney and Sun, 2000; Rahimi *et al.*, 2004; Singh *et al.*, 2006) are not devised to generate paths that directly minimize the map uncertainty. Hence, they cannot be considered greedy or non-myopic. This is also the case for the strategy of Leonard *et al.* (2007), which simplifies the path planning problem by restricting the robot observation paths to ellipses and optimizing with respect to the elliptical path parameters.

2.5 Single- vs. multi-robot strategies

A team of robots can potentially complete the task faster than a single robot. It is also more robust to failures by providing redundancy, and can reduce the hardware, energy, and payload requirements of each robot. But its performance may be adversely affected by physical interference between robots.

In contrast to single-robot exploration strategies (Batalin *et al.*, 2004; Meliou *et al.*, 2007; Popa *et al.*, 2006; Popa and Lewis, 2008; Rahimi *et al.*, 2004; Zhang and Sukhatme, 2007), our strategies have to coordinate the exploration of multiple robots like those in (Alvarez *et al.*, 2007; Leonard *et al.*, 2007; Low *et al.*, 2007; McCartney and Sun, 2000; Singh *et al.*, 2006, 2007). Using multiple robots, our proposed exploration strategies can potentially achieve lower map uncertainty than a single robot that expends the same amount of resources.

Chapter 3

Multi-Robot Adaptive Sampling Problem (MASP)

The main contribution of this chapter is to formalize the exploration problem in a sequential decision-theoretic planning under uncertainty framework called MASP. This unique formalization has a general form that unifies formulations of exploration problems with varying adaptivity. As a result, this unifying framework covers the entire adaptivity spectrum, thus subsuming various existing non-adaptive problem formulations.

We begin by formalizing the exploration problems at the two extremes of the spectrum with details of how they are to be constructed (Section 3.2). Exploration problems residing within the spectrum can be formalized in a similar manner. Readers who wish to skip the construction process can refer directly to the end of Sections 3.2.3 and 3.2.4 for the resulting problem formulations.

This unifying MASP framework is useful in allowing the performance of induced exploration policies of varying adaptivity to be theoretically analyzed and the performance advantage of a more adaptive policy to be realized (Section 3.3).

In the MASP framework, the map uncertainty is measured using the mean-squared error criterion (Section 3.2.1). Consequently, solving MASP involves planning observation

paths that produce the least expected posterior map error/uncertainty; the original MASP is therefore a cost-minimizing problem. We then provide a reward-maximizing dual formulation of MASP that, when solved, achieves the largest possible expected reduction in map error/uncertainty (Section 3.4). We show that the exploration objectives of these two problem formulations are equivalent.

The reward-maximizing MASP allows observation selection properties of the induced exploration policy to be realized for sampling the Gaussian and log-Gaussian processes (Section 3.5), namely, adaptivity, hotspot sampling, and wide-area coverage. We show that existing GP-based exploration strategies may not explore and map the hotspot field well because they are non-adaptive and do not exploit clustering phenomena (Section 3.5.1). On the other hand, the ℓGP-based exploration policy can learn a high-quality hotspot field map because it is adaptive and exploits clustering phenomena (Section 3.5.2).

3.1 Terminology and Notation

Let $\mathcal{X}$ be the domain of the hotspot field corresponding to a finite, discretized set of grid cell locations (e.g., cell centers). An observation taken (e.g., by a single robot) at stage i comprises a pair of location $x_i \in \mathcal{X}$ and its corresponding measurement z_{x_i}. More generally, k observations taken (e.g., by k robots or 1 robot taking k observations) at stage i can be represented by a pair of vectors $\mathbf{x}_i$ of k locations and $\mathbf{z}_{\mathbf{x}_i}$ of the corresponding measurements.

Definition 3.1.1 (Posterior Data). The posterior data d_i at stage $i > 0$ comprises

- the prior data $d_0 = \langle \mathbf{x}_0, \mathbf{z}_{\mathbf{x}_0} \rangle$ available at stage 0, and
- a complete history of observations $\mathbf{x}_1, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{x}_i, \mathbf{z}_{\mathbf{x}_i}$ induced by k observations per stage over stages 1 to i.

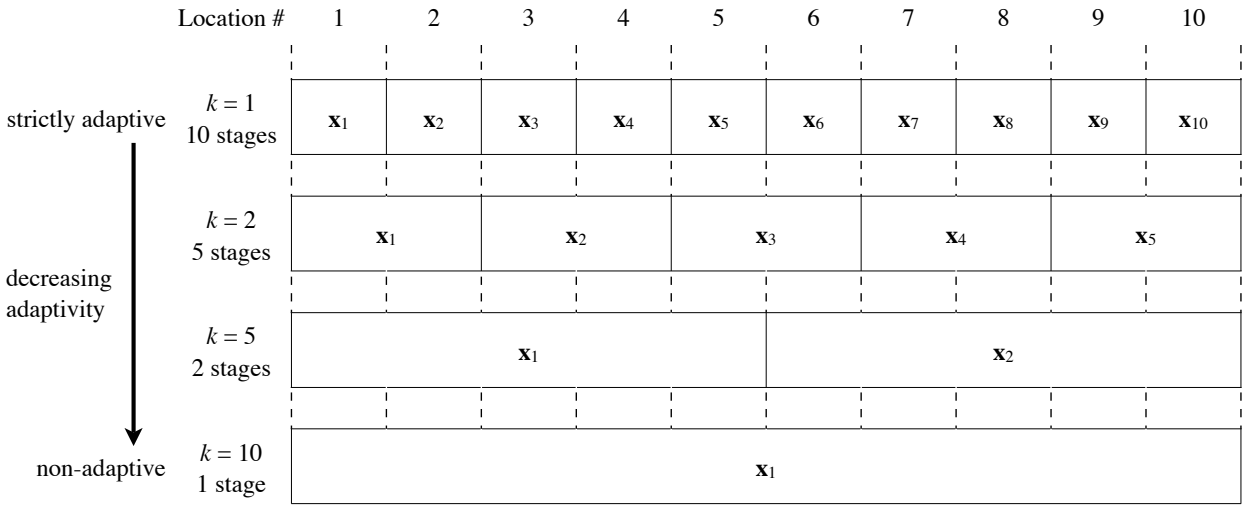


Figure 3.1: Illustration of strictly adaptive ($k = 1$), partially adaptive ($k = 2$, $k = 5$), and non-adaptive ($k = 10$) exploration strategies: the number of new locations n to be explored is 10. Each box denotes a stage consisting of the k locations aligned above it. For example, the strictly adaptive strategy spans 10 stages and selects 1 new location per stage. On the other hand, the non-adaptive strategy selects all 10 new locations in a single stage. Please refer to Definition 3.1.2 for a formal characterization of adaptive and non-adaptive exploration strategies.

Let $\mathbf{x}_{0:i}$ and $\mathbf{z}_{\mathbf{x}_{0:i}}$ denote vectors comprising the location and measurement components of the posterior data d_i (i.e., concatenations of $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_i$ and $\mathbf{z}_{\mathbf{x}_0}, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{z}_{\mathbf{x}_i}$), respectively. More generally, let $\mathbf{x}_{u:v}$ and $\mathbf{z}_{\mathbf{x}_{u:v}}$ denote, respectively, vectors concatenating $\mathbf{x}_u, \ldots, \mathbf{x}_v$ and $\mathbf{z}_{\mathbf{x}_u}, \ldots, \mathbf{z}_{\mathbf{x}_v}$ obtained from stages u to v where $0 \leq u \leq v$. Let Z_{x_i}, $\mathbf{Z}_{\mathbf{x}_i}$, $\mathbf{Z}_{\mathbf{x}_{0:i}}$, and $\mathbf{Z}_{\mathbf{x}_{u:v}}$ be the random measurements corresponding to the respective realizations z_{x_i}, $\mathbf{z}_{\mathbf{x}_i}$, $\mathbf{z}_{\mathbf{x}_{0:i}}$, and $\mathbf{z}_{\mathbf{x}_{u:v}}$.

To be more precise in our definition of adaptivity, we provide a characterization of adaptive and non-adaptive exploration strategies:

Definition 3.1.2 (Characterizing Adaptivity). Suppose that the prior data d_0 are available and n new locations are to be explored. Then, an exploration strategy is (see, for example, Fig. 3.1)

- **adaptive** if its policy to select each vector $\mathbf{x}_{i+1}$ of k new locations depends on the previously sampled data d_i for stage $i = 0, \ldots, n/k - 1$. This strategy thus selects k observations per stage over n/k stages. When $k = 1$, the strategy is strictly adaptive (Low *et al.*, 2008). Increasing k makes the strategy partially adaptive (Batalin *et al.*, 2004; Low *et al.*, 2007; Rahimi *et al.*, 2004; Singh *et al.*, 2006). When $k = n$, the strategy becomes non-adaptive as defined next;

- **non-adaptive** (Alvarez *et al.*, 2007; Leonard *et al.*, 2007; Meliou *et al.*, 2007; Popa *et al.*, 2004; Popa and Lewis, 2008; Singh *et al.*, 2007; Zhang and Sukhatme, 2007) if its policy to select each new location x_{i+1} for $i = 0, \ldots, n - 1$ is independent of the measurements $z_{x_1}, \ldots, z_{x_n}$. As a result, all n new locations $x_1, \ldots, x_n$ can be selected prior to exploration. That is, this strategy selects all n observations in a single stage.

3.2 Problem Formulations

3.2.1 Objective function

From Section 1.2, the exploration objective is to plan observation paths that minimize the uncertainty of mapping the hotspot field. To achieve this, we use the mean-squared error criterion as a measure of the map uncertainty. Supposing the posterior data d_n are available, a predictor $\hat{Z}_x(d_n)$ of the measurement z_x at the unobserved location x incurs the mean-squared error loss of

$$\mathbb{E}\{[Z_x - \hat{Z}_x(d_n)]^2 \mid d_n\} \ . \tag{3.1}$$

Then, the *posterior map error* of domain $\mathcal{X}$ with predictor $\hat{Z}_x(d_n)$ can be represented by the sum of mean-squared errors over all locations in $\mathcal{X}$, that is,

$$\sum_{x \in \mathcal{X}} \mathbb{E}\{[Z_x - \hat{Z}_x(d_n)]^2 \mid d_n\} \ . \tag{3.2}$$

Using the best unbiased predictor

$$\hat{Z}_x(d_n) \triangleq \mathbb{E}[Z_x \mid d_n]$$

(i.e., the posterior mean achieves the lowest mean-squared error among all unbiased predictors), the mean-squared error (3.1) at each location x can be reduced to the posterior variance

$$\sigma^2_{Z_x|d_n} \triangleq \text{var}[Z_x \mid d_n] \ .$$

As a result, the posterior map error (3.2) can be reduced to

$$\sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_n} \ , \tag{3.3}$$

which will be used as the minimizing criterion in the MASP formulation below.

3.2.2 Value function

If only the prior data d_0 are available, an exploration strategy has to produce a policy for selecting observation paths that minimize the *expected* posterior map error instead. This policy is therefore responsible for directing a robot team to collect the optimal observations $\mathbf{x}_1, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{x}_n, \mathbf{z}_{\mathbf{x}_n}$ during exploration to form the posterior data d_n. Given the prior data d_0, the value under an exploration policy π is defined to be the expected posterior map error

(i.e., expectation of (3.3)) when starting in d_0 and following π thereafter:

$$\begin{aligned} V_0^{\pi}(d_0) &\stackrel{\triangle}{=} \mathbb{E}\left\{\sum_{x\in\mathcal{X}} \sigma^2_{Z_x|d_n} \mid d_0, \pi\right\} \\ &= \int f(\mathbf{z}_{\mathbf{x}_{1:n}} \mid d_0, \pi) \sum_{x\in\mathcal{X}} \sigma^2_{Z_x|d_n} \, \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:n}} \end{aligned} \tag{3.4}$$

where f denotes a probability density function.

In the next two subsections, we will describe how the adaptive and non-adaptive exploration policies can be derived to minimize the expected posterior map error (3.4).

3.2.3 Adaptive exploration

The adaptive exploration policy π for directing a team of k robots is structured to collect k observations per stage over a finite planning horizon of n stages. This implies each robot observes one location per stage and is therefore constrained to explore at most n new locations over the n-stage horizon. Formally, $\pi \stackrel{\triangle}{=} \langle \pi_0(d_0), \ldots, \pi_{n-1}(d_{n-1}) \rangle$ where $\pi_i(d_i)$ maps the data state d_i to a vector of robot actions $\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)$ at stage $i = 0, \ldots, n-1$, that is,

$$\pi_i \ : \ d_i \rightarrow \mathbf{a}_i \ ,$$

and $\mathcal{A}(\mathbf{x}_i)$ is the action space of the robots (i.e., a finite set of joint actions) given their current locations $\mathbf{x}_i$. We assume that the transition function $\tau(\mathbf{x}_i, \mathbf{a}_i)$ maps the current robot locations $\mathbf{x}_i$ and actions $\mathbf{a}_i$ at stage i to the next locations $\mathbf{x}_{i+1}$ at stage $i+1$ *deterministically*, that is,

$$\tau \ : \ \mathbf{x}_i \ \times \ \mathbf{a}_i \rightarrow \mathbf{x}_{i+1} \ .$$

By putting the two functions π_i and τ together, the assignment $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i, \pi_i(d_i))$ can be obtained. We can observe from this assignment that the sequential selection of k new

locations $\mathbf{x}_{i+1}$ to be included in the observation paths depends only on the previously sampled data d_i along the paths for stage $i = 0, \ldots, n-1$. Hence, by Definition 3.1.2, the policy π is adaptive.

When the adaptive policy π is plugged into the value function of (3.4), the following n-stage recursive formulation results:

$$\begin{aligned} V_i^{\pi}(d_i) = & \int f(\mathbf{z}_{\mathbf{x}_{i+1}} \mid d_i, \pi_i)\ V_{i+1}^{\pi}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{i+1}} \\ = & \int f(\mathbf{z}_{\tau(\mathbf{x}_i, \pi_i(d_i))} \mid d_i)\ V_{i+1}^{\pi}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i, \pi_i(d_i))} \\ V_n^{\pi}(d_n) = & \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_n} \end{aligned} \tag{3.5}$$

for stage $i = 0, \ldots, n-1$. The first and second equalities follow from $f(\mathbf{z}_{\mathbf{x}_{1:n}} \mid d_0, \pi^1) = \Pi_{i=0}^{n-1} f(\mathbf{z}_{\mathbf{x}_{i+1}} \mid d_i, \pi_i^1)$ and $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i, \pi_i^1(d_i))$, respectively.

Finally, solving the adaptive exploration problem MASP(1) involves choosing the adaptive policy π to minimize $V_0^{\pi}(d_0)$ (3.5), which we call the *optimal adaptive policy* denoted by π^1. That is, π^1 is induced by the *optimal value function*

$$V_0^{\pi^1}(d_0) = \min_{\pi} V_0^{\pi}(d_0)\ . \tag{3.6}$$

By plugging π^1 into the value functions of (3.5), this optimal value function (3.6) evolves into the following n-stage dynamic programming equations:

$$\begin{aligned} V_i^{\pi^1}(d_i) = & \min_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \int f(\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \mid d_i)\ V_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \\ V_n^{\pi^1}(d_n) = & \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_n} \end{aligned} \tag{3.7}$$

for stage $i = 0, \ldots, n-1$. The optimal adaptive policy $\pi^1 = \langle \pi_0^1(d_0), \ldots, \pi_{n-1}^1(d_{n-1}) \rangle$, which

is induced by solving MASP(1), can therefore be determined in a stagewise manner by

$$\pi_i^1(d_i) = \underset{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)}{\arg\min} \int f(\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \mid d_i)\ V_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}\ . \tag{3.8}$$

From (3.8), the optimal action $\pi_0^1(d_0)$ at stage 0 can be determined prior to exploration since the prior data d_0 are available. However, each action rule $\pi_i^1(d_i)$ at stage $i = 1, \ldots, n-1$ defines the optimal action to take in response to the data d_i, part of which (i.e., $\mathbf{x}_1, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{x}_i, \mathbf{z}_{\mathbf{x}_i}$) will only be observed during exploration.

3.2.4 Non-adaptive exploration

The non-adaptive exploration policy π is structured to collect, in a single stage, n observations per robot with a team of k robots. This means each robot is also constrained to explore at most n new locations, but all of them have to do this within a single stage. Formally, $\pi \triangleq \pi_0(d_0)$ where, at stage 0, $\pi_0(d_0)$ maps the data state d_0 to a vector $\mathbf{a}_{0:n-1}$ of action components concatenating a sequence of robot actions $\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}$ ($\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)$ for $i = 0, \ldots, n-1$), that is,

$$\pi_0\ :\ d_0 \rightarrow \mathbf{a}_{0:n-1}\ .$$

By putting the two functions π_0 and τ together, the assignment $\mathbf{x}_{1:n} \leftarrow \tau(\mathbf{x}_{0:n-1}, \pi_0(d_0))$ can be obtained. We can observe from this assignment that the selection of $k \times n$ new locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to form the observation paths are independent of the measurements $\mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{z}_{\mathbf{x}_n}$ obtained along the paths during the exploration phase. Hence, by Definition 3.1.2, the policy π is non-adaptive and all new locations can be selected in a single stage prior to exploration.

When the non-adaptive policy π is plugged into the value function of (3.4), the following

single-staged formulation results:

$$\begin{aligned} V_0^{\pi}(d_0) &= \int f(\mathbf{z}_{\mathbf{x}_{1:n}} \mid d_0, \pi_0)\ V_1^{\pi}(d_n)\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:n}} \\ &= \int f(\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \pi_0(d_0))} \mid d_0)\ V_1^{\pi}(d_n)\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \pi_0(d_0))} \\ V_1^{\pi}(d_n) &= \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_n}\ . \end{aligned} \tag{3.9}$$

The second equality follows from the assignment $\mathbf{x}_{1:n} \leftarrow \tau(\mathbf{x}_{0:n-1}, \pi_0(d_0))$.

Finally, solving the non-adaptive exploration problem MASP(n) involves choosing the non-adaptive policy π to minimize $V_0^{\pi}(d_0)$ (3.9), which we call the *optimal non-adaptive policy* denoted by π^n. That is, π^n is induced by the optimal value function

$$V_0^{\pi^n}(d_0) = \min_{\pi} V_0^{\pi}(d_0)\ . \tag{3.10}$$

By plugging π^n into the value functions of (3.9), this optimal value function (3.10) evolves into the following single-staged dynamic programming equation:

$$\begin{aligned} V_0^{\pi^n}(d_0) &= \min_{\mathbf{a}_{0:n-1}} \int f(\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})} \mid d_0)\ V_1^{\pi^n}(d_n)\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})} \\ V_1^{\pi^n}(d_n) &= \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_n}\ . \end{aligned} \tag{3.11}$$

The optimal non-adaptive policy $\pi^n = \pi_0^n(d_0)$, which is induced by solving MASP(n), can therefore be determined in a single stage:

$$\pi_0^n(d_0) = \arg\min_{\mathbf{a}_{0:n-1}} \int f(\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})} \mid d_0)\ V_1^{\pi^n}(d_n)\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})}\ . \tag{3.12}$$

3.3 Performance Advantage of Adaptive Exploration

In Section 3.2, we have introduced two exploration strategies for minimizing the expected posterior map error: (i) the induced optimal adaptive policy π^1 from solving MASP(1) selects the new observations sequentially over n stages, and (ii) the induced optimal non-adaptive policy π^n from solving MASP(n) selects all new observations in a single stage. This section addresses the issue of whether the adaptive strategy performs better than the non-adaptive one.

More specifically, does the optimal adaptive policy π^1 provide a lower expected posterior map error than the optimal non-adaptive policy π^n? To know this, we can compare the induced optimal values from solving MASP(1) and MASP(n) (i.e., respectively, $V_0^{\pi^1}(d_0)$ (3.7) and $V_0^{\pi^n}(d_0)$ (3.11)), which reflect the expected posterior map error achieved by their corresponding optimal policies π^1 and π^n. It is shown in Appendix A.1 that

$$V_0^{\pi^1}(d_0) \leq V_0^{\pi^n}(d_0) \ . \tag{3.13}$$

This implies the optimal adaptive policy π^1 performs better than or at least as well as the optimal non-adaptive policy π^n in terms of the achieved expected posterior map error.

In addition, it is not computationally more expensive to solve for the optimal adaptive policy π^1 than the optimal non-adaptive policy π^n. To see this, we have to compare the amount of computation needed to solve the dynamic programming equations of MASP(1) (3.7) and MASP(n) (3.11). This can be done by keeping track of how frequently the posterior map error (i.e., $\sum_{x\in\mathcal{X}} \sigma^2_{Z_x|d_n}$) in $V_n^{\pi^1}(d_n)$ (3.7) and $V_1^{\pi^n}(d_n)$ (3.11) is evaluated. This frequency can be determined by the enumerations of different posterior data d_n, which is caused, at each stage, by

- the space of different possible actions under the minimum: if we let $\mathcal{A} = \mathcal{A}(\mathbf{x}_0) = \ldots = \mathcal{A}(\mathbf{x}_{n-1})$, the action space under the minimum is $|\mathcal{A}|$ and $|\mathcal{A}|^n$ for MASP(1) and

MASP(n), respectively;

- the space of different possible measurements under the integration: the integration for MASP(1) and MASP(n) is subject to a k- and kn-dimensional probability density function, respectively. If the integration cannot be evaluated in closed form, it has to be approximated using a numerical technique such as Monte-Carlo sampling. Suppose that we draw ν samples for the Monte-Carlo integration of MASP(1). Since the integration of MASP(n) has a probability density function of higher dimensionality, a larger number of samples (in particular, ν^n samples) has to be drawn for the numerical approximation to be effective (Press *et al.*, 2002).

Since MASP(1) spans n stages, it produces $|\mathcal{A}| \times \nu \times \ldots \times |\mathcal{A}| \times \nu = (|\mathcal{A}|\nu)^n$ enumerations of different posterior data d_n. Although MASP(n) only involves a single stage, it also produces $|\mathcal{A}|^n \nu^n$ enumerations of different posterior data. Hence, no computational advantage is gained by using the optimal non-adaptive policy π^n.

Equation 3.13 establishes the performance advantage of the optimal adaptive policy π^1 over the optimal non-adaptive policy π^n. This result can be generalized to cover the entire adaptivity spectrum. But first, a suitable adaptivity index has to be identified in order to specify exploration problems of varying adaptivity. We will use the length of action sequence per stage (i.e., under the minimum), denoted by λ, as the adaptivity index. For example, the action sequences of MASP(1) and MASP(n) per stage are of length 1 and n, respectively. Hence, a shorter action sequence per stage (i.e., smaller λ) produces a more adaptive exploration problem MASP(λ). As a result, we can form exploration problems of different adaptivity (i.e., MASP(1), ..., MASP(n)) by varying λ from 1 to n.

The generalized result (Theorem 3.3.1) indicates that in terms of the achieved expected posterior map error $V_0^{\pi^\lambda}(d_0)$, the performance of the induced optimal policy π^λ from solving MASP(λ) improves monotonically with higher adaptivity (i.e., decreasing λ):

Theorem 3.3.1. Suppose $1 \leq \lambda_1 < \lambda_2 \leq n$ and n is divisible by λ_1, λ_2. Then, $V_0^{\pi^{\lambda_1}}(d_0) \leq V_0^{\pi^{\lambda_2}}(d_0)$.

The proof of the above result is in Appendix A.1. It can be observed from the proof that, for this result to hold, it does not rely on the choice of the optimizing criterion. Hence, Theorem 3.3.1 will still be valid when we switch to the entropy criterion in Chapter 4.

3.4 Dual Formulations

In this section, we reformulate the cost-minimizing MASP(λ) as a reward-maximizing problem that lends itself to a different interpretation as described below. More importantly, the reward-maximizing problem formulation can be subject to convex analysis, which allows monotone-bounding approximations to be developed (Section 5.3).

The reward-maximizing MASP(1) comprises the following n-stage dynamic programming equations:

$$\begin{aligned} U_i^{\pi^1}(d_i) &= \max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} R^{\pi^1}(\tau(\mathbf{x}_i, \mathbf{a}_i), d_i) + \int f(\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \mid d_i)\ U_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \\ U_t^{\pi^1}(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)} R^{\pi^1}(\tau(\mathbf{x}_t, \mathbf{a}_t), d_t) \end{aligned} \tag{3.14}$$

for stage $i = 0, \ldots, t-1$ where $t = n-1$, and the reward functions $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ for $i = 0, \ldots, t$ are defined as follows:

$$\begin{aligned} R^{\pi^1}(\mathbf{x}_{i+1}, d_i) &\stackrel{\triangle}{=} \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_i} - \int f(\mathbf{z}_{\mathbf{x}_{i+1}} \mid d_i)\ \sigma^2_{Z_x|d_{i+1}}\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{i+1}} \\ &= \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_i} - \mathbb{E}[\sigma^2_{Z_x|d_{i+1}} \mid d_i] \\ &= \sum_{x \in \mathcal{X}} \mathrm{var}[\mu_{Z_x|d_{i+1}} \mid d_i] \end{aligned} \tag{3.15}$$

with $\mu_{Z_x|d_{i+1}} \triangleq \mathbb{E}[Z_x \mid d_{i+1}]$. The last equality follows from the well-known variance decomposition formula. If we consider the expression under the second equality, the stagewise reward reflects the expected map error reduction by selecting the k new locations $\mathbf{x}_{i+1}$ to be included in the observation paths. Therefore, by maximizing the sum of expected rewards in (3.14) over the n-stage horizon, the reward-maximizing MASP(1) maximizes the total expected map error reduction with the selected observation paths.

This reformulation procedure is not limited to MASP(1), but can also be applied to MASP(λ) for $1 \leq \lambda \leq n$, resulting in a similar interpretation to the above. For example, the reward-maximizing MASP(n) resolves to the following single-staged equation:

$$U_0^{\pi^n}(d_0) = \max_{\mathbf{a}_{0:n-1}} R^{\pi^n}(\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1}), d_0) \tag{3.16}$$

where

$$\begin{aligned} R^{\pi^n}(\mathbf{x}_{1:n}, d_0) &\triangleq \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_0} - \int f(\mathbf{z}_{\mathbf{x}_{1:n}} \mid d_0)\ \sigma^2_{Z_x|d_n}\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:n}} \\ &= \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_0} - \mathbb{E}[\sigma^2_{Z_x|d_n} \mid d_0]\ . \end{aligned}$$

From (3.16), the reward-maximizing MASP(n) selects $k \times n$ new locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with maximum expected map error reduction to form the observation paths.

The equivalence result below (Theorem 3.4.1) relates the cost-minimizing (3.7) and reward-maximizing (3.14) MASP(1) problem formulations in the adaptive exploration setting:

Theorem 3.4.1. The optimal value functions of the cost-minimizing (3.7) and reward-maximizing (3.14) MASP(1)'s are related by

$$V_i^{\pi^1}(d_i) = \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_i}\ - U_i^{\pi^1}(d_i) \tag{3.17}$$

for stage $i = 0, \ldots, n-1$ and their respective optimal adaptive policies coincide.

The proof of the above result is in Appendix A.2. In particular, when $i = 0$, Theorem 3.4.1 (3.17) reveals that the original exploration objective of minimizing the expected posterior map error (i.e., $V_0^{\pi^1}(d_0)$ (3.7)) is equivalent to that of applying the largest possible expected map error reduction (i.e., $U_0^{\pi^1}(d_0)$ (3.14)) to the prior map error (i.e., $\sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_0}$).

Theorem 3.4.1 can be generalized to cater to MASP(λ) for $1 \leq \lambda \leq n$ with a similar interpretation to the above. For example, the following equivalence result relates the cost-minimizing (3.11) and reward-maximizing (3.16) MASP(n) problem formulations in the non-adaptive exploration setting:

$$V_0^{\pi^n}(d_0) = \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_0} \; - U_0^{\pi^n}(d_0) \; . \tag{3.18}$$

The performance advantage of adaptive exploration can also be realized using the induced optimal values from solving the reward-maximizing MASP(λ)'s. For example, if the performance advantage of the optimal adaptive policy π^1 over the non-adaptive one π^n is quantified by the difference of their corresponding optimal values, this difference is the same for either the original or dual formulation as derived from (3.17) and (3.18):

$$V_0^{\pi^n}(d_0) - V_0^{\pi^1}(d_0) = U_0^{\pi^1}(d_0) - U_0^{\pi^n}(d_0) \; .$$

Consequently, the previously established result $V_0^{\pi^1}(d_0) \leq V_0^{\pi^n}(d_0)$ (3.13) implies

$$U_0^{\pi^1}(d_0) \geq U_0^{\pi^n}(d_0) \; .$$

This means the optimal adaptive policy π^1 achieves greater or, if not, at least equal expected map error reduction as the optimal non-adaptive policy π^n.

In the next two sections, we will show how the reward-maximizing MASP(1) (3.14) can be applied to sampling the Gaussian and log-Gaussian processes, both of which are commonly used to map environmental fields.

3.5 Learning the Hotspot Field Map

Traditionally, a hotspot is defined as a location where its measurement exceeds a pre-defined extreme. But, hotspot locations do not usually occur in isolation but in clusters. So, it is useful to characterize hotspots with spatial correlation properties. Accordingly, we define a hotspot field to vary as a realization of a spatial random field $\{Y_x > 0\}_{x\in\mathcal{X}}$ such that putting together the observed measurements of the realization $\{y_x\}_{x\in\mathcal{X}}$ gives a positively skewed 1D sample frequency distribution (e.g., Figs. 3.2a and 3.3a). In this section, we will highlight the problem with modeling the hotspot field directly using the GP and explain how the ℓGP remedies this. We will also show analytically that the MASP-based policy for sampling the ℓGP is adaptive and exploits clustering phenomena but that for sampling the GP lacks these properties.

3.5.1 Gaussian process (GP)

A widely-used random field to model environmental phenomena is the GP (Alvarez *et al.*, 2007; Meliou *et al.*, 2007; Singh *et al.*, 2007). The stationary assumption on the GP covariance structure is very sensitive to strong positive skewness of hotspot field measurements (e.g., Figs. 3.2a and 3.3a) and is easily violated by a few extreme ones (Webster and Oliver, 2007). In practice, this can cause reconstructed fields to display large hotspots centered about a few extreme observations and prediction variances to be unrealistically small in hotspots (Hohn, 1998), which are undesirable. So, if the GP is used to model a hotspot field directly, it may not map well. To remedy this, a standard statistical practice is to take

the log of the measurements (i.e., $Z_x = \log Y_x$ and $z_x = \log y_x$). Then, z_x denotes the log-measurement at each location $x \in \mathcal{X}$ with the corresponding original measurement $\exp\{z_x\}$. As shown in Fig. 3.2b, this removes a significant amount of skewness[1] and extremity from the plankton density field depicted in Fig. 1.1, resulting in an approximately symmetric 1D sample frequency distribution/histogram. As such, the GP will be used to model/map the *log-measurements* of the hotspot field instead. Consequently, the mean-squared error criterion (3.2) has to be optimized in the transformed log-scale.

We will apply MASP(1) to sampling the GP and determine if the induced exploration policy π^1 exhibits adaptive, hotspot sampling, and wide-area coverage properties. Let $\{Z_x\}_{x\in\mathcal{X}}$ denote a GP defined on the domain $\mathcal{X}$, that is, the joint distribution over any finite subset of $\{Z_x\}_{x\in\mathcal{X}}$ is Gaussian (Rasmussen and Williams, 2006). The GP can be completely specified by its (prior) mean and covariance functions

$$\mu_{Z_x} \triangleq \mathbb{E}[Z_x] \ ,$$

$$\sigma_{Z_x Z_u} \triangleq \mathrm{cov}[Z_x, Z_u]$$

for $x, u \in \mathcal{X}$. We adopt a commonly used assumption that the GP is second-order stationary (Cressie, 1993; Rasmussen and Williams, 2006). That is, it has a constant mean and a stationary covariance structure (i.e., $\sigma_{Z_x Z_u}$ is a function of $x - u$ for all $x, u \in \mathcal{X}$). If the posterior data d_n are available, the distribution of Z_x remains a Gaussian with the posterior

[1]The skewness of the measurements can be determined using the measure $m_3/(m_2\sqrt{m_2})$ (Webster and Oliver, 2007) where $m_i \triangleq |\mathcal{X}|^{-1}\sum_{x\in\mathcal{X}}(y_x - \bar{\mu})^i$ and $\bar{\mu} \triangleq |\mathcal{X}|^{-1}\sum_{x\in\mathcal{X}} y_x$. Here, we use y_x to denote the measurement at location x. A symmetric distribution has a skewness of 0. Before taking log, the original measurements of the plankton density field depicted in Fig. 1.1 show strong positive skewness of 1.447 as illustrated in Fig. 3.2b. It is recommended in (Webster and Oliver, 2007) to apply the log to the measurements when the skewness is greater than 1. After taking log, the skewness becomes -0.317, which shows a decrease in the magnitude of skewness.

mean and variance

$$\mu_{Z_x|d_n}{}^2 \stackrel{\triangle}{=} \mathbb{E}[Z_x \mid d_n] = \mu_{Z_x} + \Sigma_{x\mathbf{x}_{0:n}} \Sigma^{-1}_{\mathbf{x}_{0:n}\mathbf{x}_{0:n}} \left\{ \mathbf{z}^{\top}_{\mathbf{x}_{0:n}} - \boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{0:n}}} \right\} \tag{3.19}$$

$$\sigma^2_{Z_x|d_n} \stackrel{\triangle}{=} \mathrm{var}[Z_x \mid d_n] = \sigma_{Z_x Z_x} - \Sigma_{x\mathbf{x}_{0:n}} \Sigma^{-1}_{\mathbf{x}_{0:n}\mathbf{x}_{0:n}} \Sigma_{\mathbf{x}_{0:n}x} \tag{3.20}$$

where $\boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{0:n}}}$ is a column vector with mean components μ_{Z_v} for every location v of $\mathbf{x}_{0:n}$, $\Sigma_{x\mathbf{x}_{0:n}}$ is a covariance vector with components $\sigma_{Z_x Z_v}$ for every location v of $\mathbf{x}_{0:n}$, $\Sigma_{\mathbf{x}_{0:n}x}$ is the transpose of $\Sigma_{x\mathbf{x}_{0:n}}$, and $\Sigma_{\mathbf{x}_{0:n}\mathbf{x}_{0:n}}$ is a covariance matrix with components $\sigma_{Z_v Z_w}$ for every pair of locations v, w of $\mathbf{x}_{0:n}$. If we want to predict the original measurement at an unobserved location, say u, using the posterior data d_n, it may seem appropriate to use the best unbiased predictor, $\mu_{Z_u|d_n}$ (3.19), of the log-measurement z_u to compose the predictor $\exp\{\mu_{Z_u|d_n}\}$[3] for predicting the original measurement. This predictor is, however, biased as it underestimates the expected value of the original measurement. We will defer the description of an unbiased predictor to the next subsection (specifically, equation (3.23)). An important property of the Gaussian posterior variance $\sigma^2_{Z_x|d_n}$ (3.20) is its independence of $\mathbf{z}_{\mathbf{x}_{1:n}}$.

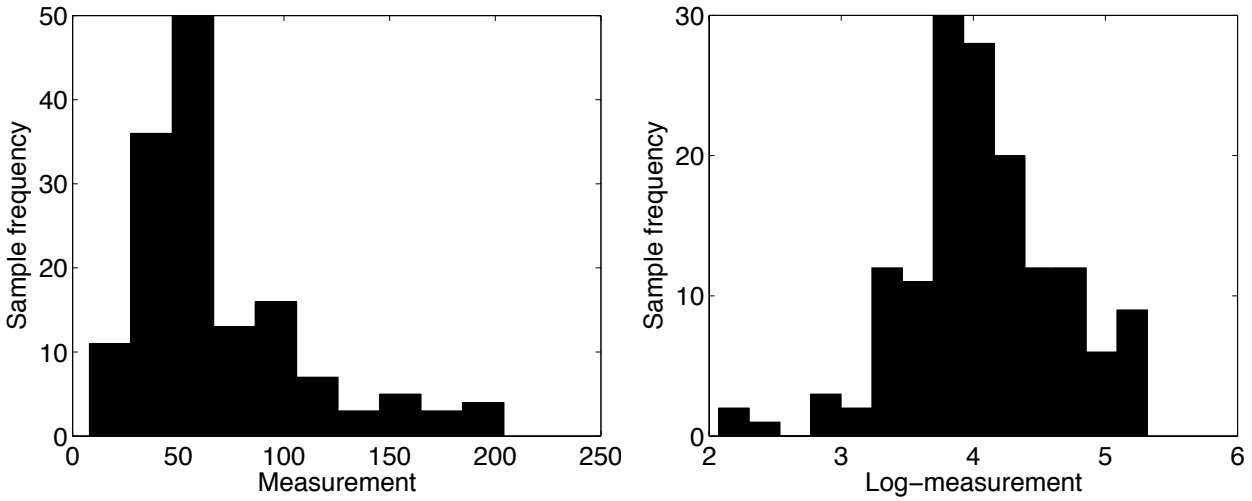
For sampling GP, the induced optimal adaptive policy π^1 from solving MASP(1) can be reduced to be *non-adaptive*. By Definition 3.1.2 of a non-adaptive policy, we have to show that the selection of new sampling locations $\mathbf{x}_{i+1}$ is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for $i = 0, \ldots, n-1$. Since the assignment $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i, \pi^1_i(d_i))$ (Section 3.2.3) reveals $\tau(\mathbf{x}_i, \pi^1_i(d_i))$ can only depend on $\mathbf{z}_{\mathbf{x}_{1:n}}$ through $\pi^1_i(d_i)$, it suffices to show that $\pi^1_i(d_i)$ is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for $i = 0, \ldots, n-1$. This is a direct consequence of the following lemma:

Lemma 3.5.1. $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ (3.15) is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for $i = 0, \ldots, n-1$.

To establish the above result, note from (3.20) that the posterior variances $\sigma^2_{Z_x|d_i}$ and

[2]Recall from Section 3.2.1 that $\mu_{Z_x|d_n}$ is the best unbiased predictor of z_x.

[3]If no posterior data d_n is available, the predictor $\exp\{\mu_{Z_u|d_n}\}$ reduces to $\exp\{\mu_{Z_u}\}$.



(a) (b)

Figure 3.2: 1D sample frequency distributions/histograms of the plankton density field (Fig. 1.1) measurements (a) before taking log, and (b) after taking log.

$\sigma^2_{Z_x|d_{i+1}}$ in the reward function $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ (3.15) are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$. The expectation in the reward function can then be integrated out to give

$$R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}} \ , \tag{3.21}$$

which is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$. For an alternative proof, please refer to Appendix A.3.

The next theorem follows from Lemma 3.5.1 and (3.14):

Theorem 3.5.1. $U_i^{\pi^1}(d_i)$ and $\pi_i^1(d_i)$ are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for $i = 0, \ldots, n-1$.

Hence, the optimal policy π^1 is *non-adaptive*. Furthermore, Theorem 3.5.1 allows the expectation in the optimal value functions of MASP(1) (3.14) to be integrated out. As a result, MASP(1) for sampling GP can be reduced to a single-staged (i.e., non-sequential)

deterministic planning problem

$$\begin{aligned} U_0^{\pi^1}(d_0) &= \sum_{i=0}^{n-1} \max_{\mathbf{a}_i} R^{\pi^1}(\tau(\mathbf{x}_i, \mathbf{a}_i), d_i) \\ &= \sum_{i=0}^{n-1} \max_{\mathbf{a}_i} \left(\sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}} \right) \\ &= \max_{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}} \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_0} - \sigma^2_{Z_x|d_n} \\ &= \max_{\mathbf{a}_{0:n-1}} R^{\pi^n}(\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1}), d_0) \\ &= U_0^{\pi^n}(d_0) \ . \end{aligned} \tag{3.22}$$

The last equality indicates that the induced optimal values from solving MASP(1) and MASP(n) are equal. So, the optimal policy π^1 does not offer any performance advantage over π^n.

For sampling GP, the induced optimal policy π^1 from solving MASP(1) performs *wide-area coverage* only. To see why, recall from (3.22) that the optimal policy π^1 selects observation paths to maximize the map error reduction (i.e., sum of variance reductions over all unobserved locations $\sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_0} - \sigma^2_{Z_x|d_n}$). If we assume isotropic covariance structure (i.e., the covariance $\sigma_{Z_x Z_u}$ decreases monotonically with $||x-u||$) (Rasmussen and Williams, 2006), the prior data d_0 provide the least amount of information on unobserved locations that are far away from the locations $\mathbf{x}_0$ observed a priori. As a result, the variances of the unobserved locations in sparsely sampled regions remain largely unreduced by the prior data d_0 collected from the observed locations $\mathbf{x}_0$. Hence, by exploring the sparsely sampled areas, a large map error reduction can be effected. Realize that the map error reduction does not account for maximization of hotspot sampling. Using the observations selected from wide-area coverage, the field of *original* measurements may not be mapped well because the under-sampled hotspots with extreme, highly-varying measurements can contribute considerably to the map error in the original scale, as discussed below.

3.5.2 Log-Gaussian process (ℓGP)

To map the original, rather than the log-, measurements directly, it is a conventional practice in geostatistics to use the non-parametric probabilistic model called the log-Gaussian process (ℓGP). Consequently, the mean-squared error criterion (3.2) is optimized in the original scale. We will show that for sampling ℓGP, π^1 is adaptive and performs both wide-area coverage and hotspot sampling.

Let $\{Y_x\}_{x\in\mathcal{X}}$ denote a ℓGP defined on the domain $\mathcal{X}$. That is, if we let $Z_x = \log Y_x$, then $\{Z_x\}_{x\in\mathcal{X}}$ is a GP (Section 3.5.1). So, the positive-valued $y_x = \exp\{z_x\}$ denotes the original measurement at each location x. The ℓGP has the (prior) mean and covariance function

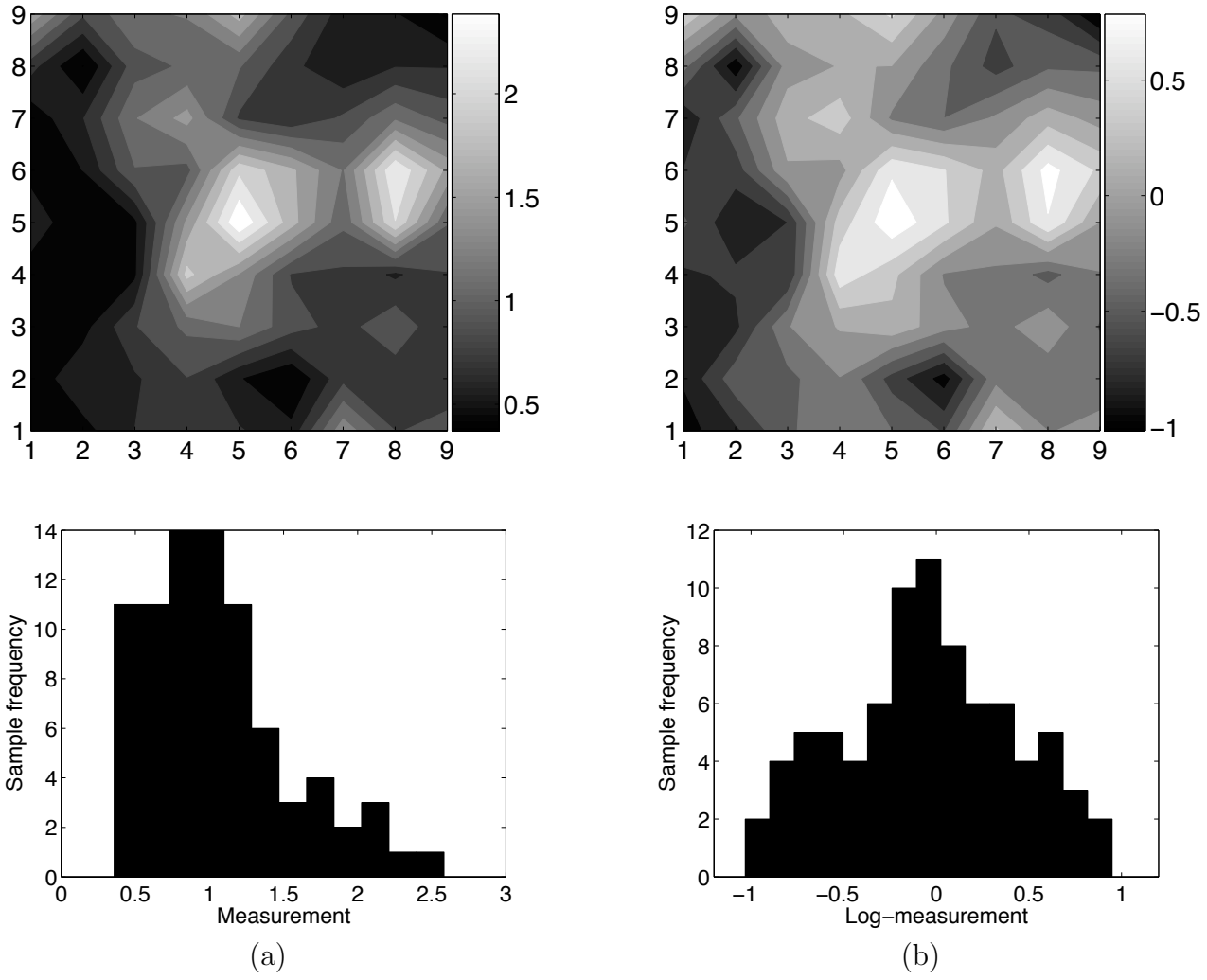
$$\mu_{Y_x} \stackrel{\triangle}{=} \mathbb{E}[Y_x] = \exp\{\mu_{Z_x} + \sigma_{Z_xZ_x}/2\} \ ,$$

$$\sigma_{Y_xY_u} \stackrel{\triangle}{=} \mathrm{cov}[Y_x, Y_u] = \mu_{Y_x}\mu_{Y_u}(\exp\{\sigma_{Z_xZ_u}\} - 1)$$

for $x, u \in \mathcal{X}$.

A ℓGP can model a field with hotspots that exhibit much higher spatial variability than the rest of the field: Figs. 3.3a and 3.3b illustrate and compare the realizations of the ℓGP and the GP; the GP realization can be obtained by taking the log of the measurements in the ℓGP realization[4]. As observed in Fig. 3.3, applying the log to the ℓGP realization has the effect of not just dampening the extreme measurements, but also dampening the difference between extreme measurements and amplifying the difference between small measurements, thus removing the positive skew (compare the 1D sample frequency distributions/histograms in Fig. 3.3). Compared to the GP realization, the ℓGP realization therefore exhibits higher

[4]To simulate a realization of the ℓGP, one has to (1) simulate a realization of the GP by drawing a random sample from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ is a vector with mean components μ_{Z_x} for $x \in \mathcal{X}$, and Σ is a covariance matrix with components $\sigma_{Z_xZ_u}$ for $x, u \in \mathcal{X}$, and (2) apply the exponential to the measurements in the GP realization. For example, the ℓGP realization in Fig. 3.3a is obtained by (1) simulating a GP realization in Fig. 3.3b with $\mathcal{X}$ being a 9×9 grid of sampling units, $\mu_{Z_x} = 0$, and $\sigma_{Z_xZ_u} = \exp\{-||x-u||^2/2(3)^2\}$ (i.e., squared exponential covariance function with a length-scale of 3) for $x, u \in \mathcal{X}$, and (2) applying the exponential to the measurements in the GP realization.



(a) (b)

Figure 3.3: Hotspot field simulation via the (a) ℓGP, and (b) GP with their respective 1D sample frequency distributions/histograms.

spatial variability within hotspots but lower spatial variability in the rest of the field. This intuitively explains why wide-area coverage suffices for the GP but hotspot sampling is further needed for the ℓGP.

From Section 3.5.1, we know that the distribution of Z_x given the posterior data d_n is Gaussian. Since the transformation from $\mathbf{z}_{\mathbf{x}_{0:n}}$ to $\mathbf{y}_{\mathbf{x}_{0:n}}$ is invertible, the distribution of Y_x given the posterior data d_n is log-Gaussian with the posterior mean and variance

$$\mu_{Y_x|d_n} \triangleq \mathbb{E}[Y_x \mid d_n] = \mathbb{E}[\exp\{Z_x\} \mid d_n] = \exp\{\mu_{Z_x|d_n} + \sigma^2_{Z_x|d_n}/2\} \tag{3.23}$$

$$\sigma^2_{Y_x|d_n} \triangleq \text{var}[Y_x \mid d_n] = \mu^2_{Y_x|d_n}(\exp\{\sigma^2_{Z_x|d_n}\} - 1) \tag{3.24}$$

where $\mu_{Z_x|d_n}$ and $\sigma^2_{Z_x|d_n}$ are the Gaussian posterior mean (3.19) and variance (3.20) respectively. Note that the log-Gaussian posterior mean $\mu_{Y_x|d_n}$ (3.23) is the best unbiased predictor for predicting the original measurement $y_x = \exp\{z_x\}$ at the unobserved location x[5].

For sampling ℓGP, the induced optimal policy π^1 from solving MASP(1) is adaptive. By Definition 3.1.2 of an adaptive policy, we have to show that the selection of new sampling locations $\mathbf{x}_{i+1}$ depends on the previously sampled data d_i for $i = 0, \ldots, n-1$. Again, it suffices to show that $\pi^1_i(d_i)$ depends on d_i for $i = 0, \ldots, n-1$. This is a direct consequence of the following lemma:

Lemma 3.5.2. $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ (3.15) depends on d_i for $i = 0, \ldots, n-1$.

To obtain the above result, it is shown in Appendix A.4 that the reward function

[5]As mentioned in the previous subsection, $\exp\{\mu_{Z_x|d_n}\}$ is a biased predictor of the original measurement $y_x = \exp\{z_x\}$. This can be observed from (3.23) that $\mu_{Y_x|d_n} \geq \exp\{\mu_{Z_x|d_n}\}$.

$R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ (3.15) can be reduced to

$$\begin{aligned} R^{\pi^1}(\mathbf{x}_{i+1}, d_i) &\triangleq \sum_{x \in \mathcal{X}} \sigma^2_{Y_x|d_i} - \mathbb{E}[\sigma^2_{Y_x|d_{i+1}} \mid d_i] \\ &= \sum_{x \in \mathcal{X}} \mu^2_{Y_x|d_i} (\exp\{\sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}\} - 1) \ . \end{aligned} \tag{3.25}$$

Since $\mu_{Y_x|d_i}$ depends on previously sampled data d_i by (3.19) and (3.23), the lemma follows. The next theorem follows from Lemma 3.5.2 and (3.14):

Theorem 3.5.2. $U_i^{\pi^1}(d_i)$ and $\pi_i^1(d_i)$ depend on d_i for $i = 0, \ldots, n-1$.

Hence, the optimal policy π^1 is *adaptive*.

For sampling ℓGP, the induced optimal adaptive policy π^1 from solving MASP(1) performs both *hotspot sampling* and *wide-area coverage*. To see this, note from the above reward function expression that the expected log-Gaussian variance reduction $\sigma^2_{Y_x|d_i} - \mathbb{E}[\sigma^2_{Y_x|d_{i+1}} \mid d_i]$ at each unobserved location x is controlled by two terms: (a) log-Gaussian posterior mean $\mu_{Y_x|d_i}$ (3.23), and (b) Gaussian variance reduction $\sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}$. A large reward can be obtained by selecting new locations $\mathbf{x}_{i+1}$ that maximize the expected log-Gaussian variance reduction at unobserved locations, in particular, those locations with (a) large log-Gaussian posterior mean (i.e., high predicted original measurement), and (b) large Gaussian variance reduction. The optimal adaptive policy π^1 therefore directs exploration towards (a) hotspots and (b) sparsely sampled areas (Section 3.5.1).

Chapter 4

Information-Theoretic Multi-Robot Adaptive Sampling Problem (iMASP)

The MASP is beset by a serious computational drawback due to its measure of map uncertainty using the mean-squared error criterion. Consequently, the time complexity of solving MASP (approximately) depends on the map resolution, which limits the practical use of MASP-based approximation algorithms in large-scale, high-resolution exploration and mapping (Chapter 5).

The principal contribution of this chapter is to alleviate this computational difficulty through an information-theoretic approach to MASP (iMASP) for efficient adaptive path planning (Section 4.1), which measures map uncertainty based on the entropy criterion (Section 4.1.1) instead. Unlike MASP, reformulating the cost-minimizing iMASP as a reward-maximizing problem (Section 4.2) causes its time complexity of being solved approximately to be independent of the map resolution and less sensitive to larger robot team size as demonstrated both analytically and empirically. In Section 4.2, we also show the equivalence between the cost-minimizing and reward-maximizing iMASPs. Beyond its computational gain, iMASP retains the beneficial properties of MASP.

Additional contributions stemming from this reward-maximizing formulation include:

- transforming the commonly-used non-adaptive maximum entropy sampling problem (Shewry and Wynn, 1987) into a novel adaptive variant, thus improving the performance of the induced exploration policy (Section 4.2);
- given an assumed environment model (e.g., occupancy grid map), establishing sufficient conditions that, when met, guarantee adaptivity provides no benefit (Section 4.3.1);
- showing analytically and empirically the superior performance of iMASP-based policies for sampling the log-Gaussian process (ℓGP) to that of policies for the widely-used Gaussian process (GP) (Guestrin *et al.*, 2005; Shewry and Wynn, 1987; Singh *et al.*, 2007) in mapping the hotspot field (Section 4.3); and
- comparing qualitatively the observation selection properties (in particular, adaptivity, hotspot sampling, and wide-area coverage) between iMASP- and MASP-based policies for sampling the GP and ℓGP (Section 4.3).

4.1 Problem Formulations

4.1.1 Objective function

From Section 1.2, the exploration objective is to plan observation paths that minimize the uncertainty of mapping the hotspot field. To achieve this, we use the entropy criterion as a measure of the map uncertainty. Let $\overline{\mathbf{x}}_{0:i}$ denote the vector comprising locations of domain $\mathcal{X}$ not observed in d_i, and $\mathbf{z}_{\overline{\mathbf{x}}_{0:i}}$ be the vector comprising the corresponding measurements. Also, let $\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}}$ be the random measurements corresponding to the realization $\mathbf{z}_{\overline{\mathbf{x}}_{0:i}}$. Supposing the posterior data d_n are available, the *posterior map entropy* of domain $\mathcal{X}$ can be represented by the posterior joint entropy of the measurements $\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}$ at the unobserved locations $\overline{\mathbf{x}}_{0:n}$:

$$\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n] \triangleq -\int f(\mathbf{z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n) \log f(\mathbf{z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n)\ d\mathbf{z}_{\overline{\mathbf{x}}_{0:n}}\ . \tag{4.1}$$

4.1.2 Value function

If only the prior data d_0 are available, an exploration strategy has to produce a policy for selecting observation paths that minimize the *expected* posterior map entropy instead. This policy must then collect the optimal observations $\mathbf{x}_1, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{x}_n, \mathbf{z}_{\mathbf{x}_n}$ during exploration to form the posterior data d_n. The value under an exploration policy π is defined to be the expected posterior map entropy (i.e., expectation of (4.1)) when starting in d_0 and following π thereafter:

$$\begin{aligned} V_0^{\pi}(d_0) &\triangleq \mathbb{E}\{\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n] \mid d_0, \pi\} \\ &= \int f(\mathbf{z}_{\mathbf{x}_{1:n}} \mid d_0, \pi)\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n]\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:n}}\ . \end{aligned} \tag{4.2}$$

The strategies of Guestrin *et al.* (2005) and Singh *et al.* (2007) have optimized a closely related *mutual information* criterion that measures the expected entropy reduction of unobserved locations $\overline{\mathbf{x}}_{0:n}$ by observing $\mathbf{x}_{1:n}$ (i.e., $\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_0] - \mathbb{E}\{\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]|d_0\}$). This is deficient for the exploration objective because mutual information may be maximized by a choice of $\mathbf{x}_{1:n}$ inducing a very large prior entropy $\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_0]$ but not necessarily the smallest expected posterior map entropy $\mathbb{E}\{\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]|d_0\}$.

In the next subsection, we will describe how the adaptive and non-adaptive exploration policies can be derived to minimize the expected posterior map entropy (4.2).

4.1.3 Adaptive and non-adaptive exploration

The process of constructing the information-theoretic exploration problems is similar to that of formulating the MASPs (Section 3.2). As shown below, the resulting cost-minimizing adaptive iMASP(1) and non-adaptive iMASP(n) differ, respectively, from MASP(1) (3.7) and MASP(n) (3.11) by only the entropy criterion. The cost-minimizing adaptive iMASP(1)

comprises the following n-stage dynamic programming equations:

$$\begin{aligned} V_i^{\pi^1}(d_i) = &\min_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \int f(\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \mid d_i)\ V_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \\ V_n^{\pi^1}(d_n) = &\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n] \end{aligned} \tag{4.3}$$

for stage $i = 0, \ldots, n-1$.

On the other hand, the cost-minimizing non-adaptive iMASP(n) evolves into the following single-staged equation:

$$\begin{aligned} V_0^{\pi^n}(d_0) = &\min_{\mathbf{a}_{0:n-1}} \int f(\mathbf{z}_{\tau(\mathbf{x}_{0:n-1},\mathbf{a}_{0:n-1})} \mid d_0)\ V_1^{\pi^n}(d_n)\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_{0:n-1},\mathbf{a}_{0:n-1})} \\ V_1^{\pi^n}(d_n) = &\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n]\ . \end{aligned} \tag{4.4}$$

The optimal adaptive and non-adaptive policies are also similar to that of MASP(1) (3.8) and MASP(n) (3.12), respectively.

4.2 Dual Formulations

In this section, we transform the cost-minimizing iMASP(1) (4.3) and iMASP(n) (4.4) into reward-maximizing problems and show their equivalence. As we shall see below, though the reward-maximizing iMASP(1) and iMASP(n) differ, respectively, from MASP(1) (3.14) and MASP(n) (3.16) by only the entropy-based reward functions, the reward-maximizing iMASPs become significantly different from the MASPs in terms of interpretation and computational complexity.

The reward-maximizing iMASP(n) turns out to be the well-known *maximum entropy sampling* (MES) problem (Shewry and Wynn, 1987):

$$U_0^{\pi^n}(d_0) = \max_{\mathbf{a}_{0:n-1}} R^{\pi^n}(\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1}), d_0) \tag{4.5}$$

where

$$R^{\pi^n}(\mathbf{x}_{1:n}, d_0) \triangleq \mathbb{H}[\mathbf{Z}_{\mathbf{x}_{1:n}} \mid d_0] \ ,$$

which is a single-staged problem of selecting $k \times n$ new locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with maximum entropy to form the observation paths. This dual ensues from the equivalence result

$$V_0^{\pi^n}(d_0) = \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_0} \mid d_0] - U_0^{\pi^n}(d_0)$$

relating cost-minimizing and reward-maximizing iMASP(n)'s in the non-adaptive exploration setting, which follows from the chain rule of entropy. This result says the original objective of minimizing expected posterior map entropy (i.e., $V_0^{\pi^n}(d_0)$ (4.4)) is equivalent to that of discharging from prior map entropy $\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_0} \mid d_0]$ the largest possible entropy into the selected paths (i.e., $U_0^{\pi^n}(d_0)$ (4.5)). Hence, their optimal non-adaptive policies coincide.

Our reward-maximizing iMASP(1) is a novel adaptive variant of MES. Unlike the cost-minimizing iMASP(1), it can be subject to convex analysis, which allows monotone-bounding approximations to be developed (Section 5.3). It comprises the following n-stage dynamic programming equations:

$$\begin{aligned} U_i^{\pi^1}(d_i) &= \max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} R^{\pi^1}(\tau(\mathbf{x}_i, \mathbf{a}_i), d_i) + \int f(\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \mid d_i)\ U_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \\ U_t^{\pi^1}(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)} R^{\pi^1}(\tau(\mathbf{x}_t, \mathbf{a}_t), d_t) \end{aligned} \tag{4.6}$$

for stage $i = 0, \ldots, t-1$ where $t = n-1$, and the reward functions $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ for $i = 0, \ldots, t$ are defined as follows:

$$R^{\pi^1}(\mathbf{x}_{i+1}, d_i) \triangleq \mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i] \ . \tag{4.7}$$

Each stagewise reward reflects the entropy of k new locations $\mathbf{x}_{i+1}$ to be potentially selected into the paths. By maximizing the sum of expected rewards over n stages in (4.6), the reward-maximizing iMASP(1) absorbs the largest expected entropy into the selected paths.

In the adaptive exploration setting, the cost-minimizing and reward-maximizing iMASP(1)'s are also equivalent (i.e., their optimal adaptive policies coincide):

Theorem 4.2.1. The optimal value functions of the cost-minimizing (4.3) and reward-maximizing (4.6) iMASP(1)'s are related by

$$V_i^{\pi^1}(d_i) = \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}} \mid d_i] - U_i^{\pi^1}(d_i)$$

for stage $i = 0, \ldots, n-1$ and their respective optimal adaptive policies coincide.

Theorem 3.4.1 has also provided an equivalence result to relate the cost-minimizing and reward-maximizing MASP(1)'s through the use of the variance decomposition formula in its induction proof. In contrast, the induction proof to Theorem 4.2.1 (see Appendix A.5) uses the chain rule of entropy, which entails a computational complexity reduction (not available to MASP(1)) as described next.

In cost-minimizing iMASP(1), the time complexity of evaluating the cost (i.e., posterior map entropy (4.1)) depends on the domain size $|\mathcal{X}|$ for the environment models described in the next subsection. By transforming into the dual, the time complexity of evaluating each stagewise reward (4.7) becomes independent of $|\mathcal{X}|$ because it reflects only the uncertainty of the new locations to be potentially selected into the observation paths. As a result, the runtime of the approximation algorithm proposed in Chapter 5 does not depend on the map resolution, which is clearly advantageous in large-scale, high-resolution exploration and mapping. In contrast, the reward-maximizing MASP(1) (3.14) utilizing the mean-squared error criterion does not share this computational advantage, as the time needed to evaluate each stagewise reward (3.15) still depends on $|\mathcal{X}|$. We will evaluate this computational advantage using time complexity analysis in Section 5.5.

4.3 Learning the Hotspot Field Map

4.3.1 Gaussian process (GP)

We will apply iMASP(1) to sampling the GP and determine if the induced exploration policy π^1 exhibits adaptive, hotspot sampling, and wide-area coverage properties.

For sampling GP, the induced optimal adaptive policy π^1 from solving iMASP(1) can be reduced to be *non-adaptive*: observe from Appendix A.6 that each stagewise reward $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ (4.7) is independent of the measurements

$$R^{\pi^1}(\mathbf{x}_{i+1}, d_i) \stackrel{\triangle}{=} \mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i] = \log \sqrt{(2\pi e)^k \ |\Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}|} \tag{4.8}$$

where $\Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}$ is a covariance matrix with components $\sigma_{Z_x Z_u|d_i}$ (i.e., for every pair of locations x, u of $\mathbf{x}_{i+1}$) that are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$. Hence, Lemma 3.5.1 holds. As a result, it follows from (4.6) that $U_i^{\pi^1}(d_i)$ and $\pi_i^1(d_i)$ are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for $i = 0, \ldots, n-1$. So, Theorem 3.5.1 also holds. The expectations in iMASP(1) (3.14) can then be integrated out. As a result, iMASP(1) for sampling GP can be reduced to a single-staged deterministic planning problem

$$\begin{aligned}
U_0^{\pi^1}(d_0) &= \sum_{i=0}^{n-1} \max_{\mathbf{a}_i} R^{\pi^1}(\tau(\mathbf{x}_i, \mathbf{a}_i), d_i) \\
&= \sum_{i=0}^{n-1} \max_{\mathbf{a}_i} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \mid d_i] \\
&= \max_{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}} \sum_{i=0}^{n-1} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \mid d_i] \\
&= \max_{\mathbf{a}_{0:n-1}} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})} \mid d_0] \\
&= \max_{\mathbf{a}_{0:n-1}} R^{\pi^n}(\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1}), d_0) \\
&= U_0^{\pi^n}(d_0) \ .
\end{aligned} \tag{4.9}$$

This indicates the induced optimal values from solving iMASP(1) and iMASP(n) are equal.

So, π^1 offers no performance advantage over π^n.

Based on the analyses of (4.9) above and (3.22) in Section 3.5.1, the following sufficient conditions, when met, guarantee that adaptivity has no benefit under an assumed environmental model:

Theorem 4.3.1. If $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for stage $i = 0, \ldots, n-1$, iMASP(1) and π^1 can be reduced to be single-staged and non-adaptive, respectively.

For example, Theorem 4.3.1 also holds for the simple case of an *occupancy grid map* modeling an obstacle-ridden environment, which typically assumes z_x for $x \in \mathcal{X}$ to be independent. As a result, $R^{\pi^1}(\mathbf{x}_{i+1}, d_i) \triangleq \mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i]$ can be reduced to a sum of prior entropies over the unobserved locations $\mathbf{x}_{i+1}$, which are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$.

For sampling GP, the induced optimal policy π^1 from solving iMASP(1) performs *wide-area coverage* only: to maximize stagewise rewards (4.8), policy π^1 selects new locations with large posterior variance for observation. If we assume isotropic covariance structure (i.e., the covariance $\sigma_{Z_x Z_u}$ decreases monotonically with $||x-u||$) (Rasmussen and Williams, 2006), the posterior data d_i provide the least amount of information on unobserved locations that are far away from the observed locations $\mathbf{x}_i$. As a result, the variances of unobserved locations in sparsely sampled regions are still largely unreduced by the posterior data d_i collected from the observed locations $\mathbf{x}_i$. Hence, by exploring the sparsely sampled areas, a large expected entropy can be absorbed into the selected observation paths.

Recall from Section 3.5.1 that, for sampling GP, the induced optimal policy π^1 from solving MASP(1) also displays non-adaptive and wide-area coverage properties. Compared to the iMASP(1)-based policy, it is expected to explore areas that are more sparsely sampled (i.e., better wide-area coverage): (a) while the iMASP(1)-based policy only considers the variances of locations to be visited by its selected paths (4.8), the MASP(1)-based policy

further considers how much its selected paths reduce the variances of unobserved locations (3.21); (b) the iMASP(1)-based policy penalizes locations of extremely high variances (i.e., by the logarithm in the reward function expression (4.8)), which typically reside in more sparsely-sampled areas. Consequently, the MASP(1)-based policy is capable of exploring regions that are more sparsely sampled. These also intuitively explain why the MASP(1)-based policy can achieve better mapping in terms of the mean-squared error criterion. However, we expect the wide-area coverage behavior and mapping performance (i.e., in the mean-squared error sense) of the MASP(1)-based policy to be more similar to that of iMASP(1)-based policy as the length-scale hyperparameter of the GP decreases.

Using the observations selected from wide-area coverage, the field of *original* measurements may not be mapped well because the under-sampled hotspots with extreme, highly-varying measurements contribute considerably to map entropy in the original scale, as discussed below.

4.3.2 Log-Gaussian process (ℓGP)

For sampling ℓGP, the induced optimal policy π^1 from solving MASP(1) is *adaptive*: observe from Appendix A.7 that each stagewise reward $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ (4.7) depends on the previously sampled data d_i:

$$R^{\pi^1}(\mathbf{x}_{i+1}, d_i) \triangleq \mathbb{H}[\mathbf{Y}_{\mathbf{x}_{i+1}} \mid d_i] = \log \sqrt{(2\pi e)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mathbf{1}^\top \tag{4.10}$$

where $\boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}$ is a mean vector with components $\mu_{Z_x|d_i}$ for every location x of $\mathbf{x}_{i+1}$. Since $\mu_{Z_x|d_i}$ depends on d_i by (3.19), $\mathbb{H}[\mathbf{Y}_{\mathbf{x}_{i+1}} \mid d_i]$ depends on d_i. So, Lemma 3.5.2 holds. Consequently, it follows from (3.14) that $U_i^{\pi^1}(d_i)$ and $\pi_i^1(d_i)$ depend on d_i for $i = 0, \ldots, n-1$. Hence, Theorem 3.5.2 holds, thus indicating that the optimal policy π^1 is *adaptive*.

For sampling ℓGP, the induced optimal adaptive policy π^1 from solving MASP(1) per-

forms both *hotspot sampling* and *wide-area coverage*: to maximize the stagewise rewards (4.10), π^1 selects new locations with (a) large Gaussian posterior variance and (b) large Gaussian posterior mean for observation. So, it directs exploration towards (a) sparsely sampled areas (Section 4.3.1) and (b) hotspots.

Recall from Section 3.5.2 that, for sampling ℓGP, the induced optimal policy π^1 from solving MASP(1) also displays adaptive, hotspot sampling, and wide-area coverage properties. Using a similar reasoning as that in Section 4.3.1, it is expected to explore areas that are more sparsely sampled (i.e., better wide-area coverage) than the iMASP(1)-based policy. The MASP(1)-based policy is also expected to sample hotspots with more extreme, highly-varying measurements (i.e., better hotspot sampling) as it favors locations of high predicted original measurements (i.e., large log-Gaussian posterior mean[1]) (3.25). In contrast, the iMASP(1)-based policy only considers locations of high predicted log-measurements (i.e., large Gaussian posterior mean) (4.10). Unlike the iMASP(1)-based policy, the MASP(1)-based policy takes into account the predicted measurements of unobserved locations (3.25), thus enabling it to sample potentially wider hotspots. Though the iMASP(1)-based policy is less effective in these observation selection properties, it bears a considerable computational gain over the MASP(1)-based policy as described previously in Section 4.2 and shown analytically in Section 5.5. We shall see later in Section 6 that, on two real-world datasets, the iMASP-based policy can empirically achieve mapping performance comparable to the MASP-based policy using significantly less time, and its incurred planning time is also less sensitive to larger robot team size. This makes the iMASP-based planner more practical for real-time deployment.

[1]Recall from (3.23) that a large log-Gaussian posterior mean can be achieved by large Gaussian posterior mean and variance.

Chapter 5

Value-Function Approximations

In this chapter, we will exploit the problem structure of strictly adaptive MASP and iMASP (Section 5.1) for sampling the ℓGP to derive approximately optimal exploration policies in a computationally tractable manner. To handle continuous states, the convexity of reward-maximizing MASP and iMASP allows discrete-state monotone-bounding approximations to be developed (Section 5.3). Consequently, we can provide theoretical guarantees on the performance of approximately optimal vs. optimal adaptive policies (Section 5.3), and establish theoretical bounds quantifying the performance advantage of optimal adaptive over non-adaptive policies (Section 5.4). We then propose anytime algorithms (Section 5.5) based on the approximate MASP and iMASP to alleviate the computational difficulty that arises from their non-Markovian structure. As demonstrated analytically in Section 5.5.2, the time complexity of the iMASP-based anytime algorithm is independent of map resolution and less sensitive to increasing robot team size as compared to the MASP-based algorithm.

5.1 Strictly Adaptive Exploration

We have described in Sections 3.4 and 4.2 how the optimal adaptive policy π^1 can be produced by solving the reward-maximizing MASP(1) (3.15) (iMASP(1) (4.7)). However, if the

robot team has more than 1 robot (i.e., $k > 1$), this optimal policy π^1 becomes *partially adaptive* because it collects more than 1 observation (i.e., k observations) per stage (see Definition 3.1.2). In this section, we will focus on deriving the optimal *strictly adaptive* policy (in particular, for sampling ℓGP), which, among policies of all adaptivity, (a) achieves the largest expected map error reduction when the mean-squared error criterion (Section 3.2.1) is being optimized or (b) absorbs the largest expected entropy into observation paths when the entropy criterion (Section 4.1.1) is being optimized.

By Definition 3.1.2, a strictly adaptive policy has to be structured to collect only 1 observation per stage. The reward-maximizing MASP(1) (3.14) (iMASP(1) (4.6)) can be revised in the following ways to impose strict adaptivity:

1. The space $\mathcal{A}(\mathbf{x}_i)$ of simultaneous joint actions is reduced to a constrained set $\mathcal{A}'(\mathbf{x}_i)$ of joint actions that, in each stage i, allows one robot to move to observe a new location and the other robots stay put. This tradeoff for strict adaptivity allows the constrained action set $\mathcal{A}'(\mathbf{x}_i)$ to grow linearly, rather than exponentially, with the number of robots;

2. We constrain each robot to explore a path of at most n new adjacent locations; this can be viewed as an energy consumption constraint on each robot. The planning horizon then spans $k \times n$ stages, rather than n, stages, which reflects the additional time of exploration incurred by strict adaptivity;

3. If the robot actions $\mathbf{a}_i$ can only be chosen from the constrained action set $\mathcal{A}'(\mathbf{x}_i)$, the assignment $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i, \mathbf{a}_i)$ moves one chosen robot to a new location x_{i+1} while the other unselected robots stay put at their current locations. Then, only one component in the current robot locations $\mathbf{x}_i$ is changed to the new location x_{i+1} to form the next locations $\mathbf{x}_{i+1}$; the other components in $\mathbf{x}_{i+1}$ are unchanged from $\mathbf{x}_i$. Formally, x_{i+1} is the component in $\mathbf{x}_{i+1}$ with the same index as the only non-zero component in $\mathbf{x}_{i+1} - \mathbf{x}_i$. Hence, there is only one unobserved random component $Y_{x_{i+1}}$ in the random

measurements $\mathbf{Y}_{\mathbf{x}_{i+1}}$; the other components in $\mathbf{Y}_{\mathbf{x}_{i+1}}$ have already been observed in the previous stages and can be found in the known data d_i. As a result, the probability distribution of the random measurements $\mathbf{Y}_{\mathbf{x}_{i+1}}$ can be simplified to a uni-variate random measurement $Y_{x_{i+1}}$, which reduces the computational burden of solving the problem numerically.

These revisions of MASP(1) (iMASP(1)) yield the strictly adaptive exploration problem called MASP($\frac{1}{k}$)[1] (iMASP($\frac{1}{k}$)). Since the MASP(1) and iMASP(1) share a common reward-maximizing problem structure, the MASP($\frac{1}{k}$) and iMASP($\frac{1}{k}$) also share a similar problem structure consisting of the following optimal value functions:

$$\begin{aligned} U_i(d_i) &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \int f(y_{x_{i+1}} \mid d_i)\ U_{i+1}(d_{i+1})\ \mathrm{d}y_{x_{i+1}} \\ &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \mathbb{E}[U_{i+1}(d_i, x_{i+1}, Y_{x_{i+1}}) \mid d_i] \\ U_t(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t) \end{aligned} \tag{5.1}$$

for stage $i = 0, \ldots, t-1$ where $t = kn - 1$. For stage $i = 0, \ldots, t$, the reward functions $R(x_{i+1}, d_i)$ of MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) are defined in a similar manner as the reward functions $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ of the reward-maximizing MASP(1) (3.15) (iMASP(1) (4.7)). Without ambiguity, we have omitted the superscript $\pi^{\frac{1}{k}}$ from the reward and value functions above. The optimal strictly adaptive policy $\pi^{\frac{1}{k}} = \langle \pi_0^{\frac{1}{k}}(d_0), \ldots, \pi_t^{\frac{1}{k}}(d_t) \rangle$ is produced by solving MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)).

Since the random measurement $Y_{x_{i+1}}$ is continuous, it entails an infinite number of state transitions. So, the conditional expectation $\mathbb{E}[U_{i+1}(d_i, x_{i+1}, Y_{x_{i+1}}) \mid d_i]$ for stage $i = 0, \ldots, t-$

[1]Recall from Section 3.3 that the length of action sequence per stage is used as the adaptivity index. However, the action sequences of MASP(1) and MASP($\frac{1}{k}$) are of the same length. Hence, we need an adaptivity index of finer resolution but consistent with the existing index being used. We will use the relative batch size per stage (i.e., number of observations per stage weighted by $\frac{1}{k}$) as the adaptivity index. Since a strictly adaptive policy collects 1 observation per stage, the adaptivity index of the strictly adaptive exploration problem is $\frac{1}{k}$. On the other hand, the optimal partially adaptive policy of MASP(1) collects k observations per stage and still evaluates to the adaptivity index of 1.

1 has to be evaluated in closed form for MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) to be solved exactly. This can be performed for MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) with $t = 1$, which can consequently be solved exactly (Section 8.2.1). At this moment, we are not aware of any computationally feasible methods to solve MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) with $t > 1$ exactly because the expectation of the optimal value function results in an integral that is too complex to be evaluated. Hence, we will resort to approximating MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) as described below. For ease of exposition, we will revert to using the $Z_{x_{i+1}} = \log Y_{x_{i+1}}$ variable for the ℓGP in the rest of this chapter.

5.2 Related Work on Sequential Decision-Theoretic Planning with Continuous States

The difficulty of solving MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) lies exactly in evaluating the conditional expectation with respect to the continuous-state random measurement $Y_{x_{i+1}}$. This intricate issue of handling continuous states is faced by the following related classes of sequential decision-theoretic planning problems, which have resolved it by constructing approximate problems:

1. Markov decision processes (MDPs). The conventional approach of generalizing to continuous states in an MDP is to approximate the value function with a parameterized model; the resulting solution is usually hard to analyze and may diverge (Bertsekas and Tsitsiklis, 1996; Sutton, 1998). Another commonly used technique is to approximate the conditional expectation using Monte-Carlo sampling (Rust, 1997), which suffers from an exponential blow-up in the state size with an increasing number of stages.

   The issue of generalizing to continuous states has been a recent focus of time-dependent MDPs (Boyan and Littman, 2001; Feng *et al.*, 2004; Li and Littman, 2005; Marecki *et al.*, 2007), and factored MDPs (Guestrin *et al.*, 2004; Hauskrecht and Kveton, 2004;

Table 5.1: Comparison of structural constraints on time-dependent MDPs with respect to the continuous state.

MDP \ Function	Transition Function	Reward Function	Value Function
(Boyan and Littman, 2001)	discrete	piecewise-linear	piecewise-linear
(Feng *et al.*, 2004)	discrete	piecewise-constant/linear	piecewise-constant/linear
(Li and Littman, 2005)	piecewise-constant	piecewise-constant	piecewise-constant
(Marecki *et al.*, 2007)	exponential	constant	piecewise-gamma

Kveton and Hauskrecht, 2006; Kveton *et al.*, 2006). A time-dependent MDP is generalized by augmenting its discrete state components with a continuous time component. To make it computationally feasible to solve, it has to be approximated by constraining its transition, reward, and value functions to certain function families as shown in Table 5.1. For the time-dependent MDPs in (Boyan and Littman, 2001; Feng *et al.*, 2004; Li and Littman, 2005), the approximation can be improved by refining the discretization or increasing the number of piecewise functions. However, this will result in an exponential blow-up with an increasing number of stages, which restricts applicability to problems of only a few stages. Although the time-dependent MDP in (Marecki *et al.*, 2007) can be solved in closed form, the number of breakpoints to be numerically determined can grow exponentially, which entails an exponential number of disjoint intervals and value function evaluations for these intervals. Furthermore, the structural assumptions are extremely restrictive: (a) similar to the MDP in (Boyan and Littman, 2001), the transition function (i.e., probability distribution of the time component) does not depend on any continuous state in the previous stages, which simplifies the identification of breakpoints, (b) if the transition function does not adhere to the exponential distribution, an approximation error results, but is not captured in the policy performance guarantee, and (c) the reward does not depend on the continuous time component.

In a factored MDP, the value function is approximated by a linear combination of basis functions and optimized via linear programming. For the factored MDPs in (Guestrin *et al.*, 2004; Hauskrecht and Kveton, 2004; Kveton and Hauskrecht, 2006), the continuous state is bounded on the $[0, 1]$ interval, the transition function is restricted to a mixture of beta distributions, and the basis functions can be piecewise-linear, polynomials or beta distributions. The factored MDP in (Kveton *et al.*, 2006) restricts the transition and basis functions to exponential-family distributions. These restrictions on the transition and basis functions allow the conditional expectation to be evaluated to a closed-form solution. A serious drawback with this approach is that a continuous-state MDP will induce infinitely many constraints in the linear programming formulation. As a result, the constraint space has to be approximated (e.g., through Monte-Carlo sampling of constraints (Hauskrecht and Kveton, 2004)), which reduces the policy performance.

In contrast to MDPs, MASP and *i*MASP adopt a more complex but realistic non-Markovian structure: the state transitions and rewards are conditioned on the entire history of actions and continuous states. More importantly, by assuming the reward and value functions to be convex, piecewise-linear functions can be constructed to monotonically lower- and upper-bound the value function (Section 5.3). Note that the form of transition function is not restricted.

2. Non-Markov problems. Bayes sequential design problems (Brockwell and Kadane, 2003; Müller *et al.*, 2007) and stochastic programs (Birge and Wets, 1986; Casey and Sen, 2005; Dupačová *et al.*, 2000; Edirisinghe, 1999; Frauendorfer, 1996; Frauendorfer and Haarbrücker, 2003; Frauendorfer and Schürle, 2000; Shapiro, 2006) can be modeled as non-Markov dynamic programming problems. In contrast to MASP and *i*MASP, they have a simple structure: (a) their transition functions do not depend on past actions, (b) for Bayes sequential design, the entire history of continuous states can be reduced

to a summary (not necessarily sufficient) statistic, and (c) for stochastic programs, the reward function is often assumed to be linear in the action variable (Birge and Wets, 1986; Casey and Sen, 2005; Edirisinghe, 1999; Frauendorfer and Haarbrücker, 2003). To make them computationally feasible to solve, the conditional expectation is approximated using Monte-Carlo sampling for both problems (Brockwell and Kadane, 2003; Dupačová *et al.*, 2000; Müller *et al.*, 2007; Shapiro, 2006) and bounding methods (Birge and Wets, 1986; Casey and Sen, 2005; Edirisinghe, 1999; Frauendorfer, 1996; Frauendorfer and Haarbrücker, 2003; Frauendorfer and Schürle, 2000) for stochastic programs. The latter technique further assumes the value function to be linear or convex in the continuous state and action variables. The resulting approximate problems suffer from an exponential blow-up in the state size as the number of stages increases. Our proposed bounding approximation technique (Section 5.3) utilizes the results on generalized Jensen and Edmundsen-Madansky bounds for convex functions (Huang *et al.*, 1977) from the field of stochastic programming.

5.3 Approximately Optimal Exploration

Our technique of approximating MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) is to construct approximate problems from the original problem MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)), and use the induced optimal policy from solving an approximate problem as an approximately optimal policy in the original problem. This section describes how the approximate problems are constructed and then provides a theoretical guarantee on the policy quality for use in the original problem.

Recall from Sections 5.1 and 5.2 that the difficulty of solving MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) lies exactly in evaluating the conditional expectation $\mathbb{E}[U_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}}) \mid d_i]$ (5.1) with respect to the continuous-state random measurement $Z_{x_{i+1}}$ for stage $i = 0, \ldots, t-1$, which cannot be evaluated in closed form. So, to approximate MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)), we should

examine how the conditional expectation can be approximated. To do this, we claim that the optimal value function $U_{i+1}(d_i, x_{i+1}, z_{x_{i+1}})$ is convex in $z_{x_{i+1}}$, which will be proven later in Lemma 5.3.1.

Let the support of $Z_{x_{i+1}}$ given the posterior data d_i be $\mathcal{Z}^{\nu}_{x_{i+1}}$ that is partitioned into ν non-empty, disjoint intervals $\mathcal{Z}^{[j]}_{x_{i+1}} \triangleq [\overline{z}^{[j-1]}_{x_{i+1}}, \overline{z}^{[j]}_{x_{i+1}}]$ for $j = 1, \ldots, \nu$. Then, the conditional expectation can be approximated from below and above using the ν-fold generalized Jensen and Edmundsen-Madansky (EM) bounds (Huang *et al.*, 1977), respectively, that is,

$$\sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} U_{i+1}(d_i, x_{i+1}, \underline{z}^{[j]}_{x_{i+1}}) \leq \mathbb{E}[U_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}}) \mid d_i] \leq \sum_{j=0}^{\nu} \overline{p}^{[j]}_{x_{i+1}} U_{i+1}(d_i, x_{i+1}, \overline{z}^{[j]}_{x_{i+1}}) \tag{5.2}$$

where

$$\underline{p}^{[j]}_{x_{i+1}} \triangleq \int_{\mathcal{Z}^{[j]}_{x_{i+1}}} f(z_{x_{i+1}}|d_i)\ \mathrm{d}z_{x_{i+1}} \qquad \text{and} \qquad \underline{z}^{[j]}_{x_{i+1}} \triangleq \frac{1}{\underline{p}^{[j]}_{x_{i+1}}} \int_{\mathcal{Z}^{[j]}_{x_{i+1}}} z_{x_{i+1}}\ f(z_{x_{i+1}}|d_i)\ \mathrm{d}z_{x_{i+1}}$$

for $j = 1, \ldots, \nu$,

$$\overline{p}^{[j]}_{x_{i+1}} \triangleq \underline{p}^{[j]}_{x_{i+1}} \left(\frac{\underline{z}^{[j]}_{x_{i+1}} - \overline{z}^{[j-1]}_{x_{i+1}}}{\overline{z}^{[j]}_{x_{i+1}} - \overline{z}^{[j-1]}_{x_{i+1}}} \right) + \underline{p}^{[j+1]}_{x_{i+1}} \left(\frac{\overline{z}^{[j+1]}_{x_{i+1}} - \underline{z}^{[j+1]}_{x_{i+1}}}{\overline{z}^{[j+1]}_{x_{i+1}} - \overline{z}^{[j]}_{x_{i+1}}} \right)$$

for $j = 0, \ldots, \nu$, and $\underline{p}^{[0]}_{x_{i+1}} := \underline{p}^{[\nu+1]}_{x_{i+1}} := \underline{z}^{[0]}_{x_{i+1}} := \underline{z}^{[\nu+1]}_{x_{i+1}} := \overline{z}^{[-1]}_{x_{i+1}} := 0$. For example, when $\nu = 1$, the 1-fold generalized Jensen bound (5.2) reduces to the Jensen's inequality (5.3), that is, $\underline{p}^{[1]}_{x_{i+1}} = 1$ and $\underline{z}^{[1]}_{x_{i+1}} = \mathbb{E}[Z_{x_{i+1}} \mid d_i]$. Then, the conditional expectation can be approximated from below using Jensen's inequality:

$$U_{i+1}(d_i, x_{i+1}, \mathbb{E}[Z_{x_{i+1}} \mid d_i]) \leq \mathbb{E}[U_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}}) \mid d_i]\ . \tag{5.3}$$

By increasing ν to refine the partition, the generalized Jensen and EM bounds can be improved. In Appendix A.8, we show that the generalized Jensen and EM bounds (5.2) can

be obtained by constructing piecewise-linear functions to lower- and upper-bound a convex function, respectively. The generalized Jensen bounds can also be viewed as approximating the continuous state variable $Z_{x_{i+1}}$ using a discrete one with a distribution at points $\underline{z}^{[j]}_{x_{i+1}}$ of probability $\underline{p}^{[j]}_{x_{i+1}} > 0$ for $j = 1, \ldots, \nu$ where $\sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} = 1$. This is similarly true for the generalized EM bounds.

To construct the *lower approximate* strictly adaptive exploration problem denoted by $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$), the conditional expectation in $\text{MASP}(\frac{1}{k})$ ($i\text{MASP}(\frac{1}{k})$) (5.1) is replaced by the lower ν-fold generalized Jensen bound (5.2). This yields the following optimal value functions of $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$):

$$\begin{aligned} \underline{U}^{\nu}_{i}(d_i) &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} \underline{U}^{\nu}_{i+1}(d_i, x_{i+1}, \underline{z}^{[j]}_{x_{i+1}}) \\ \underline{U}^{\nu}_{t}(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t) \end{aligned} \tag{5.4}$$

for stage $i = 0, \ldots, t-1$ where $\underline{p}^{[j]}_{x_{i+1}}$ and $\underline{z}^{[j]}_{x_{i+1}}$ correspond to those of the generalized Jensen bound (5.2). The induced optimal policy $\underline{\pi}^{\frac{1}{k}} = \langle \underline{\pi}^{\frac{1}{k}}_0(d_0), \ldots, \underline{\pi}^{\frac{1}{k}}_t(d_t) \rangle$ is produced by solving $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$) and is used as an approximately optimal policy in the original problem. In the same way, the *upper approximate* problem denoted by $\overline{\text{MASP}}(\frac{1}{k})$ ($\overline{i\text{MASP}}(\frac{1}{k})$) can be constructed from $\text{MASP}(\frac{1}{k})$ ($i\text{MASP}(\frac{1}{k})$) (5.1) by replacing the conditional expectation with the upper ν-fold generalized EM bound (5.2), which results in a problem structure similar to that of (5.4) with the optimal value functions $\overline{U}^{\nu}_{i}(d_i)$ for $i = 0, \ldots, t$.

We will now show that (a) when the mean-squared error criterion (Section 3.2.1) is being optimized, the largest expected map error reduction $U_0(d_0)$ achieved by the optimal strictly adaptive policy $\pi^{\frac{1}{k}}$ can be approximated from below and above using the induced optimal values $\underline{U}^{\nu}_{0}(d_0)$ and $\overline{U}^{\nu}_{i}(d_i)$ from solving $\underline{\text{MASP}}(\frac{1}{k})$ (5.4) and $\overline{\text{MASP}}(\frac{1}{k})$, respectively; and (b) when the entropy criterion (Section 4.1.1) is being optimized, the largest expected entropy of observation paths $U_0(d_0)$ achieved by policy $\pi^{\frac{1}{k}}$ can be approximated from below and

above using the induced optimal values $\underline{U}_0^{\nu}(d_0)$ and $\overline{U}_i^{\nu}(d_i)$ from solving $\underline{i\text{MASP}}(\frac{1}{k})$ (5.4) and $\overline{i\text{MASP}}(\frac{1}{k})$, respectively. To do this, we make use of a stronger convexity result for the optimal value functions of MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)):

Lemma 5.3.1. $U_i(d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ for $i = 0, \ldots, t$.

The proof of Lemma 5.3.1 is provided in Appendix A.9.

The following result uses the induced optimal values $\underline{U}_0^{\nu}(d_0)$ and $\overline{U}_i^{\nu}(d_i)$ from, respectively, solving the lower and upper approximate problems to monotonically bound the (a) largest expected map error reduction $U_0(d_0)$ achieved by the induced optimal strictly adaptive policy $\pi^{\frac{1}{k}}$ from solving MASP($\frac{1}{k}$); and the (b) largest expected entropy of observation paths $U_0(d_0)$ achieved by policy $\pi^{\frac{1}{k}}$ from solving iMASP($\frac{1}{k}$). By increasing ν to refine the partition, these bounds can be improved. But, this increases the computational burden of solving the approximate problems.

Theorem 5.3.1. If $\mathcal{Z}_{x_{i+1}}^{\nu+1}$ is obtained by splitting one of the intervals in $\mathcal{Z}_{x_{i+1}}^{\nu}$, $\underline{U}_i^{\nu}(d_i) \leq \underline{U}_i^{\nu+1}(d_i) \leq U_i(d_i) \leq \overline{U}_i^{\nu+1}(d_i) \leq \overline{U}_i^{\nu}(d_i)$ for $i = 0, \ldots, t$.

The proof of Theorem 5.3.1 is provided in Appendix A.10.

The next result provides pessimistic estimates of the (a) minimum expected posterior map error $V_0(d_0)$ (i.e., induced optimal value from solving the cost-minimizing MASP($\frac{1}{k}$)); and the (b) minimum expected posterior map entropy $V_0(d_0)$ (i.e., induced optimal value from solving the cost-minimizing iMASP($\frac{1}{k}$)). It follows directly from Theorems 3.4.1, 4.2.1, and 5.3.1:

Corollary 5.3.1. For MASP($\frac{1}{k}$) and iMASP($\frac{1}{k}$), $V_0(d_0) \leq \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_0} - \underline{U}^{\nu}_0(d_0)$ and $V_0(d_0) \leq \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}} \mid d_i] - \underline{U}^{\nu}_0(d_0)$, respectively.

The following result tells us how the approximately optimal policy $\underline{\pi}^{\frac{1}{k}}$ performs against the optimal strictly adaptive policy $\pi^{\frac{1}{k}}$ in terms of the achieved expected map error reduction (expected entropy of observation paths). Clearly, it does not perform better than policy $\pi^{\frac{1}{k}}$. But, we guarantee that policy $\underline{\pi}^{\frac{1}{k}}$ can achieve an expected map error reduction (expected entropy of observation paths) not worse than $\underline{U}^{\nu}_i(d_i)$:

Theorem 5.3.2. Define the expected map error reduction (expected entropy of observation paths) achieved by an adaptive exploration policy π with the following value functions

$$\begin{aligned} U^{\pi}_i(d_i) = &\ R(\tau(\mathbf{x}_i, \pi_i(d_i)), d_i) + \int f(\mathbf{z}_{\tau(\mathbf{x}_i,\pi_i(d_i))} \mid d_i)\ U^{\pi}_{i+1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i,\pi_i(d_i))} \\ U^{\pi}_t(d_t) = &\ R(\tau(\mathbf{x}_t, \pi_t(d_t)), d_t) \end{aligned}$$

for stage $i = 0, \ldots, t-1$. Then, $\underline{U}^{\nu}_i(d_i) \leq U^{\underline{\pi}^{\frac{1}{k}}}_i(d_i) \leq U_i(d_i)$ for stage $i = 0, \ldots, t$.

The proof of Theorem 5.3.2 is provided in Appendix A.11.

The above result does not account for how much the expected map error reduction $U^{\underline{\pi}^{\frac{1}{k}}}_0(d_0)$ (expected entropy of observation paths) achieved by the policy $\underline{\pi}^{\frac{1}{k}}$ differs from that (i.e., $U_0(d_0)$) achieved by the optimal policy $\pi^{\frac{1}{k}}$. With the upper bound of Theorem 5.3.1, this error difference $U_0(d_0) - U^{\underline{\pi}^{\frac{1}{k}}}_0(d_0)$ can be bounded:

Corollary 5.3.2. Let $\epsilon \triangleq \overline{U}_0^{\nu}(d_0) - \underline{U}_0^{\nu}(d_0)$. Then, policy $\underline{\pi}^{\frac{1}{k}}$ is ϵ-optimal for achieving the mean-squared error (entropy) criterion. That is, $U_0(d_0) - U_0^{\underline{\pi}^{\frac{1}{k}}}(d_0) \leq \epsilon$.

In other words, policy $\underline{\pi}^{\frac{1}{k}}$ is guaranteed to achieve an expected map error reduction $U_0^{\underline{\pi}^{\frac{1}{k}}}(d_0)$ (expected entropy of observation paths) that is not more than $\overline{U}_0^{\nu}(d_0) - \underline{U}_0^{\nu}(d_0)$ from the largest expected map error reduction (expected entropy of observation paths) $U_0(d_0)$ achieved by $\pi^{\frac{1}{k}}$.

5.4 Bounds on Performance Advantage of Adaptive Exploration

Previously, we have established the performance advantage of optimal adaptive over non-adaptive policies (Section 3.3). Realizing the extent of such an advantage is important if adaptivity incurs a cost (e.g., additional time of exploration incurred by strict adaptivity in Section 5.1). In particular, we are interested in quantifying the performance difference between the strictly adaptive $\pi^{\frac{1}{k}}$ and the non-adaptive π^n. This performance advantage of $\pi^{\frac{1}{k}}$ over π^n is defined as the difference of their achieved expected map error reduction or expected entropy of observation paths $U_0(d_0) - U_0^{\pi^n}(d_0)$. Using the induced optimal values from solving the approximate problems (Theorem 5.3.1), the advantage $U_0(d_0) - U_0^{\pi^n}(d_0)$ can be bounded between $\underline{U}_0^{\nu}(d_0) - U_0^{\pi^n}(d_0)$ and $\overline{U}_0^{\nu}(d_0) - U_0^{\pi^n}(d_0)$. A large lower bound $\underline{U}_0^{\nu}(d_0) - U_0^{\pi^n}(d_0)$ implies that $\pi^{\frac{1}{k}}$ is to be preferred. A small upper bound $\overline{U}_0^{\nu}(d_0) - U_0^{\pi^n}(d_0)$ implies that π^n performs close to that of $\pi^{\frac{1}{k}}$ and should be preferred if it is more costly to deploy $\pi^{\frac{1}{k}}$. For GP, note that this advantage is zero because $\pi^{\frac{1}{k}}$ can be reduced to be non-adaptive as shown previously.

5.5 Real-Time Dynamic Programming

For the bounding approximation scheme in Section 5.3, the state size grows exponentially with the number of stages. This is due to the nature of dynamic programming problems (e.g., $\underline{\text{MASP}}(\frac{1}{k})$ ($i\underline{\text{MASP}}(\frac{1}{k})$) (5.4)), which take into account all possible states. To alleviate this computational difficulty, we propose an anytime algorithm based on $\underline{\text{MASP}}(\frac{1}{k})$ ($i\underline{\text{MASP}}(\frac{1}{k})$), which can guarantee their policy performance in real time.

Our proposed anytime algorithm is adapted from the *Real-Time Dynamic Programming* (RTDP) (Barto *et al.*, 1995) technique, which is a well-known heuristic search algorithm for discrete-state MDPs. RTDP essentially simulates greedy exploration paths through a large state space. This results in the following desirable properties: (a) the search is focused, that is, it does not have to evaluate the entire state space to obtain the optimal policy, and (b) it has a good anytime behavior, that is, it produces a good policy fast and this policy improves over time. The disadvantage of RTDP is its slow convergence due to the focused search.

A non-trivial issue arises with generalizing RTDP to handle the non-Markovian problem structure of $\underline{\text{MASP}}(\frac{1}{k})$ ($i\underline{\text{MASP}}(\frac{1}{k})$) (5.4): the state space of the MDP is often assumed to be tractable. Based on this assumption, the RTDP has been enhanced in (Bonet and Geffner, 2003a,b) with additional procedures to improve convergence whose computation time is linear in the state size. More importantly, improvements of RTDP (Bonet and Geffner, 2003a,b; McMahan *et al.*, 2005; Smith and Simmons, 2006) emphasize the use of informed heuristic bounds, which are preprocessed with time complexity linear in the state size. This is clearly unacceptable for our anytime algorithms, since the state size of $\underline{\text{MASP}}(\frac{1}{k})$ ($i\underline{\text{MASP}}(\frac{1}{k})$) (5.4) grows exponentially with the number of stages. In the next subsection, we will derive informed heuristic bounds that are computationally efficient.

5.5.1 Preprocessing of heuristic bounds

The greedy exploration in RTDP is guided by heuristic bounds, which are used to prune unnecessary, bad searches of the state space while still guaranteeing policy optimality. In particular, when the initial bounds are more informed or tighter (as opposed to non-informed, loose bounds used in (Barto *et al.*, 1995)), the anytime and convergence performance can be improved (Bonet and Geffner, 2003a,b; McMahan *et al.*, 2005). However, this usually entails a higher computational complexity in processing the bounds. For example, in a reward-maximizing MDP with discrete states, the greedy search is guided by admissible upper bounds, which can be obtained through deterministic relaxation of the problem (Bonet and Geffner, 2003a,b; McMahan *et al.*, 2005; Smith and Simmons, 2006). One such form of relaxation is to choose the best possible outcome/next state of any action, which is illustrated below for $\underline{\text{MASP}}(\frac{1}{k})$ ($i\underline{\text{MASP}}(\frac{1}{k})$) (5.4):

$$\begin{aligned} \underline{U}_i^{\nu}(d_i) &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \max_{j \in \{1,\ldots,\nu\}} \underline{U}_{i+1}^{\nu}(d_i, x_{i+1}, z_{x_{i+1}}^{[j]}) \\ \underline{U}_t^{\nu}(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t) \end{aligned} \tag{5.5}$$

for stage $i = 0, \ldots, t-1$. Clearly, the induced optimal values from solving the deterministic relaxation (5.5) upper-bound that of $\underline{\text{MASP}}(\frac{1}{k})$ ($i\underline{\text{MASP}}(\frac{1}{k})$) (5.4) over all stages. The state space (i.e., the set of all possible data episodes), however, grows exponentially with the number of stages t (i.e., $O(\nu^t \Pi_{i=0}^t |\mathcal{A}'(\mathbf{x}_i)|)$), thus implying exponential time complexity to enumerate all data episodes for computing the upper bound. In the paragraphs below, we will derive informed initial bounds that are computationally efficient.

Before doing so, it is noteworthy to point out that similar to the improved RTDP techniques in (McMahan *et al.*, 2005; Smith and Simmons, 2006), our anytime algorithms maintain both lower and upper bounds, and the approximately optimal policy is induced from the lower bounds. This is in contrast to RTDP (Barto *et al.*, 1995) and its other variants

(Bonet and Geffner, 2003a,b), which maintain and use only upper bounds to generate the action policy. By keeping two-sided bounds for each encountered state, the uncertainty of its corresponding optimal value function can be derived and exploited to guide future searches in a more informed manner. More importantly, this design choice accounts for the real-time guarantee on the policy performance.

To derive informed initial lower bounds, we can try to perform a deterministic relaxation of MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) (5.1). By assuming the optimal value functions of MASP($\frac{1}{k}$) (iMASP($\frac{1}{k}$)) to be convex (Lemma 5.3.1), this relaxation can be achieved through $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$) (5.4) with $\nu = 1$:

$$\begin{aligned} \underline{U}_i^1(d_i) = &\max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \underline{U}_{i+1}^1(d_i, x_{i+1}, \mathbb{E}[Z_{x_{i+1}} \mid d_i]) \\ \underline{U}_t^1(d_t) = &\max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t) \end{aligned} \tag{5.6}$$

for stage $i = 0, \ldots, t-1$. As shown in Theorem 5.3.1, the induced optimal values from solving the deterministic relaxation (5.6) lower-bound that of $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$) (5.4) over all stages. Note that the state size still grows exponentially with the number of stages t (i.e., $O(|\mathcal{A}'|^t)$ if $\mathcal{A}'(\mathbf{x}_i) = \mathcal{A}'$ for $i = 0, \ldots, t$).

To obtain computationally efficient informed lower bounds, (5.6) can be relaxed further by choosing the best action to maximize the immediate reward at each stage:

$$\begin{aligned} \underline{H}_i(d_i) = &\; R(x_{i+1}^*, d_i) + \underline{H}_{i+1}(d_i, x_{i+1}^*, \mathbb{E}[Z_{x_{i+1}^*} \mid d_i]) \\ \underline{H}_t(d_t) = &\; R(x_{t+1}^*, d_t) \end{aligned} \tag{5.7}$$

for $i = 0, \ldots, t-1$ where $R(x_{i+1}^*, d_i) = \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i)$. It is easy to see that $\underline{H}_i(d_i) \leq \underline{U}_i^1(d_i)$ and therefore lower-bounds the induced optimal values from solving $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$)

(5.4). It is shown in Appendix A.12 that

$$\underline{H}_i(d_i) \leq \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}_{x_{i+1}}^{[j]} \underline{H}_{i+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]})$$

with $\underline{p}_{x_{i+1}}^{[j]}$ and $\underline{z}_{x_{i+1}}^{[j]}$ defined according to that of $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$) (5.4). In this case, we say that the lower heuristic bound is monotonic. However, the state size only grows linearly with the number of stages t (i.e., $O(t)$), thus requiring linear time complexity to enumerate all data episodes for computing the lower bound.

To derive computationally efficient informed upper bounds for $\underline{\text{MASP}}(\frac{1}{k})$ (5.4), Theorem 3.4.1 can be exploited to give

$$\begin{aligned} \overline{H}_i(d_i) &= \textstyle\sum_{x \in \mathcal{X}} \sigma^2_{Y_x|d_i} \\ \overline{H}_t(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t) \end{aligned} \tag{5.8}$$

for stage $i = 0, \ldots, t-1$. It can be observed from Theorem 3.4.1 that $\overline{H}_i(d_i)$ upper-bounds $U_i(d_i)$ for stage $i = 0, \ldots, t$ since $V_i(d_i) \geq 0$. Therefore, they upper-bound the induced optimal values from solving $\underline{\text{MASP}}(\frac{1}{k})$. The time complexity of evaluating this upper bound is constant in the number of stages. It is shown in Appendix A.13 that

$$\overline{H}_i(d_i) \geq \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}_{x_{i+1}}^{[j]} \overline{H}_{i+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]})$$

with $\underline{p}_{x_{i+1}}^{[j]}$ and $\underline{z}_{x_{i+1}}^{[j]}$ defined according to that of $\underline{\text{MASP}}(\frac{1}{k})$ (5.4). In this case, we say that the upper heuristic bound is monotonic.

For $\underline{i\text{MASP}}(\frac{1}{k})$ (5.4), since $V_i(d_i)$ may be negative, we cannot do likewise. If it can be

assumed that $V_i(d_i) \geq 0$, Theorem 4.2.1 can then be exploited to give

$$\begin{aligned} \overline{H}_i(d_i) &= \mathbb{H}[\mathbf{Y}_{\overline{\mathbf{x}}_{0:i}} \mid d_i] \\ \overline{H}_t(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t) \end{aligned} \tag{5.9}$$

for stage $i = 0, \ldots, t-1$ such that $\overline{H}_i(d_i)$ upper-bounds $U_i(d_i)$ for stage $i = 0, \ldots, t$. Therefore, they upper-bound the induced optimal values from solving $\underline{i\text{MASP}}(\frac{1}{k})$.

5.5.2 Anytime algorithms

The anytime algorithm (Algorithm 1) is inspired by uncertainty-based RTDP (URTDP) techniques (McMahan *et al.*, 2005; Smith and Simmons, 2006). To elaborate, each simulated exploration path involves an alternating selection of actions and their corresponding outcomes until the last stage is reached. Each action is selected based on the upper bound (line 3). For each encountered state, the algorithm maintains both lower and upper bounds, which are used to derive the uncertainty of its corresponding optimal value function. It exploits them to guide future searches in an informed manner: it explores the next state/outcome with the greatest amount of uncertainty (lines 4-5). Then, the algorithm backtracks up the path to update the upper heuristic bounds using $\max_{\mathbf{a}_i} \overline{Q}_i(\mathbf{a}_i, d_i)$ (line 12) where

$$\overline{Q}_i(\mathbf{a}_i, d_i) \triangleq R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} \overline{U}_{i+1}(d_i, x_{i+1}, \underline{z}^{[j]}_{x_{i+1}})$$

and the lower bounds via $\max_{\mathbf{a}_i} \underline{Q}_i(\mathbf{a}_i, d_i)$ (line 13) where

$$\underline{Q}_i(\mathbf{a}_i, d_i) \triangleq R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} \underline{U}_{i+1}(d_i, x_{i+1}, \underline{z}^{[j]}_{x_{i+1}}) \ .$$

We assume that whenever a new state is encountered, it is initialized with the computationally efficient informed lower and upper bounds derived in Section 5.5.1. When an exploration policy is requested at any time during the algorithm's execution, we provide the greedy policy that is induced by the lower bound. The policy performance has a similar guarantee to Theorem 5.3.2: the URTDP algorithm for solving $\underline{\text{MASP}}(\frac{1}{k})$ ($\underline{i\text{MASP}}(\frac{1}{k})$) provides a greedy policy that can achieve an expected map error reduction (expected entropy of observation paths) not worse than $\underline{U}_0(d_0)$.

URTDP(d_0, t):
  while $\overline{U}_0(d_0) - \underline{U}_0(d_0) > \alpha$ **do** SIMULATED-PATH(d_0, t)

SIMULATED-PATH(d_0, t):
1: $i \leftarrow 0$
2: **while** $i < t$ **do**
3: $\quad \mathbf{a}_i^* \leftarrow \arg\max_{\mathbf{a}_i} \overline{Q}_i(\mathbf{a}_i, d_i)$
4: $\quad \forall j, \;\; \Xi_j \leftarrow \underline{p}^{[j]}_{x^*_{i+1}} \{\overline{U}_{i+1}(d_i, x^*_{i+1}, \underline{z}^{[j]}_{x^*_{i+1}}) - \underline{U}_{i+1}(d_i, x^*_{i+1}, \underline{z}^{[j]}_{x^*_{i+1}})\}$
5: $\quad z \leftarrow$ sample from distribution at points $\underline{z}^{[j]}_{x^*_{i+1}}$ of probability $\Xi_j / \sum_k \Xi_k$
6: $\quad d_{i+1} \leftarrow d_i, x^*_{i+1}, z$
7: $\quad i \leftarrow i + 1$
8: $\overline{U}_i(d_i) \leftarrow \max_{\mathbf{a}_i} R(x_{i+1}, d_i)$
9: $\underline{U}_i(d_i) \leftarrow \max_{\mathbf{a}_i} R(x_{i+1}, d_i)$
10: **while** $i > 0$ **do**
11: $\quad i \leftarrow i - 1$
12: $\quad \overline{U}_i(d_i) \leftarrow \max_{\mathbf{a}_i} \overline{Q}_i(\mathbf{a}_i, d_i)$
13: $\quad \underline{U}_i(d_i) \leftarrow \max_{\mathbf{a}_i} \underline{Q}_i(\mathbf{a}_i, d_i)$

Algorithm 1: URTDP (α is user-specified bound).

We will show that the time complexity of running the SIMULATED-PATH(d_0, t) procedure is independent of map resolution for the $\underline{i\text{MASP}}(\frac{1}{k})$-based URTDP algorithm but the time complexity of running the same procedure for the $\underline{\text{MASP}}(\frac{1}{k})$-based URTDP algorithm is not. It is also less sensitive to increasing robot team size. Assuming no prior data and $\mathcal{A}' = \mathcal{A}'(\mathbf{x}_0) = \ldots = \mathcal{A}'(\mathbf{x}_t)$, the time needed to evaluate the stagewise rewards

$R(x_{i+1}, d_i) \stackrel{\triangle}{=} \mathbb{H}[Y_{x_{i+1}} \mid d_i]$ for all $|\mathcal{A}'|$ new locations x_{i+1} (i.e., using Cholesky factorization) is $\mathcal{O}(t^3 + |\mathcal{A}'|t^2)$, which is independent of $|\mathcal{X}|$ and results in $\mathcal{O}(t(t^3 + |\mathcal{A}'|(t^2 + \nu)))$ time to run the SIMULATED-PATH(d_0, t) procedure for the $\underline{i\text{MASP}}(\frac{1}{k})$-based URTDP algorithm. In contrast, the time needed to evaluate the stagewise rewards $R(x_{i+1}, d_i) \stackrel{\triangle}{=} \sum_{x \in \mathcal{X}} \mu^2_{Y_x|d_i}(\exp\{\sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}\} - 1)$ is $\mathcal{O}(t^3 + |\mathcal{A}'|(t^2 + |\mathcal{X}|t) + |\mathcal{X}|t^2)$, which depends on $|\mathcal{X}|$ and entails $\mathcal{O}(t(t^3 + |\mathcal{A}'|(t^2 + |\mathcal{X}|t + \nu) + |\mathcal{X}|t^2))$ time to run the same procedure for the $\underline{\text{MASP}}(\frac{1}{k})$-based URTDP algorithm. When the joint action set size $|\mathcal{A}'|$ increases with larger robot team size, the time to run the SIMULATED-PATH(d_0, t) procedure for the $\underline{\text{MASP}}(\frac{1}{k})$-based URTDP algorithm increases faster than that for the $\underline{i\text{MASP}}(\frac{1}{k})$-based URTDP algorithm due to the gradient factor $|\mathcal{X}|t$ involving large domain size. In the next chapter, we will report the time taken to run this procedure empirically.

Chapter 6

Experiments and Discussion

This chapter evaluates, empirically, the induced approximately optimal strictly adaptive policy $\underline{\pi}^{\frac{1}{k}}$ from solving $\underline{i\text{MASP}}(\frac{1}{k})$ on 2 real-world datasets exhibiting positive skew: (a) June 2006 plankton density data (Fig. 6.1a) of Chesapeake Bay bounded within latitude $38.481 - 38.591$N and longitude $76.487 - 76.335$W, and (b) potassium distribution data (Fig. 6.1d) of Broom's Barn farm spanning 520m by 440m. Each region is discretized into a 14×12 grid of sampling units. Each unit x is, respectively, associated with (a) plankton density y_x (chl-a) in mg m^{-3}, and (b) potassium level y_x (K) in mg l^{-1}. Each region comprises, respectively, (a) $|\mathcal{X}| = 148$ and (b) $|\mathcal{X}| = 156$ such units. Using a team of 2 robots, each robot is tasked to explore 9 adjacent units in its path including its starting unit. If only 1 robot is used, it is placed, respectively, in (a) top and (b) bottom starting unit, and samples all 18 units. Each robot's actions are restricted to move to the front, left, or right unit. We use the data of 20 randomly selected units to learn the hyperparameters (i.e., mean and covariance structure) of GP and ℓGP through maximum likelihood estimation (Rasmussen and Williams, 2006). So, the prior data d_0 comprise the randomly selected and robot starting units.

The performance of $\underline{\pi}^{\frac{1}{k}}$ is compared to the policies produced by four state-of-the-art exploration strategies: The *optimal non-adaptive policy* π^n *for GP* (Shewry and Wynn,

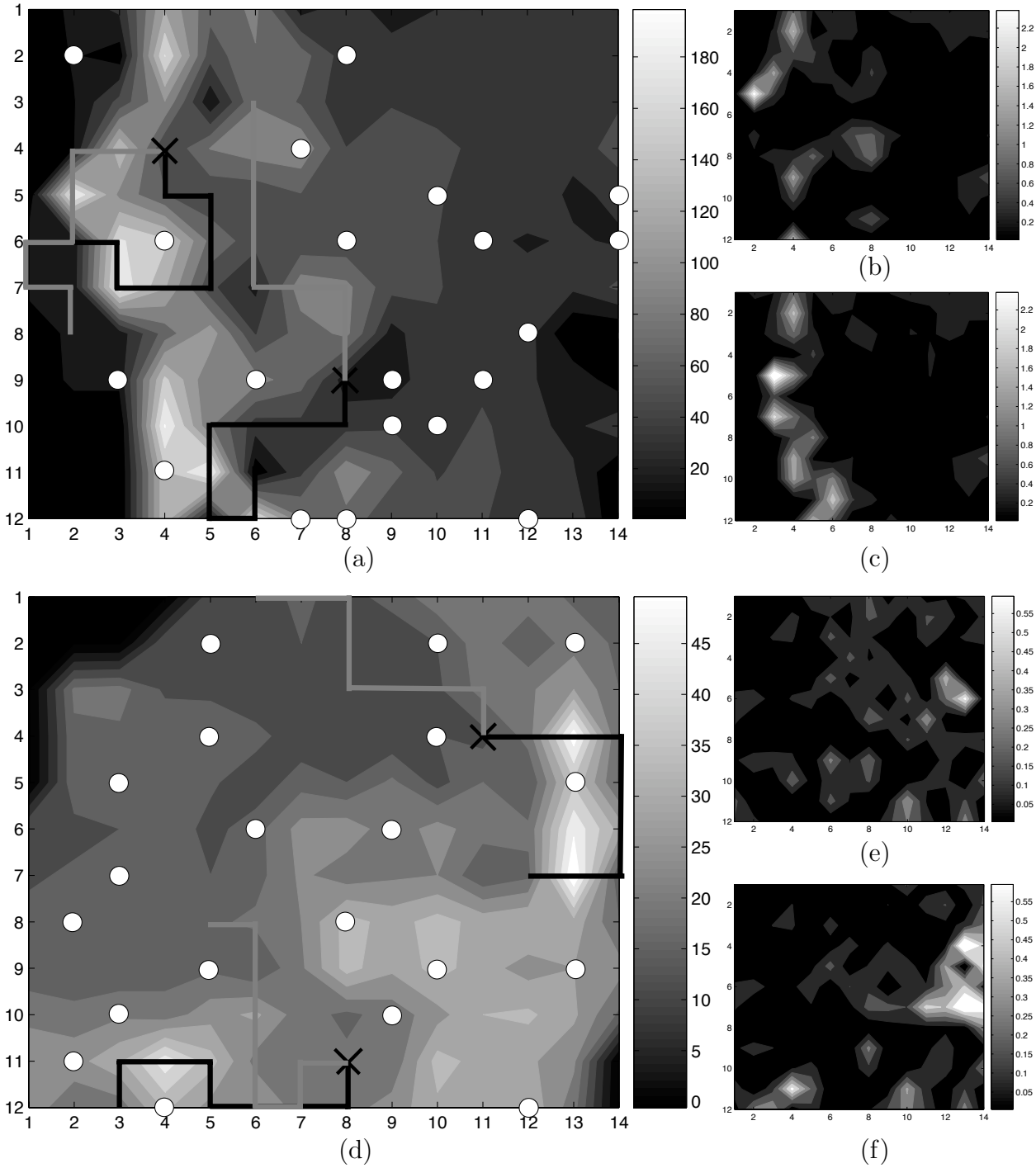


Figure 6.1: (a) chl-a field with prediction error maps for (b) strictly adaptive $\underline{\pi}^{\frac{1}{k}}$ and (c) non-adaptive π^n: 20 units (white circles) are randomly selected as prior data. The robots start at locations marked by '$\times$'s. The black and gray robot paths are produced by $\underline{\pi}^{\frac{1}{k}}$ and π^n respectively. (d-f) K field with prediction error maps for $\underline{\pi}^{\frac{1}{k}}$ and π^n.

1987) is produced by solving iMASP(n) (3.16). Similar to Theorem 3.4.1, it can be shown to be equivalent to the strictly adaptive $\pi^{\frac{1}{k}}$ for GP. Although $\underline{i\text{MASP}}(\frac{1}{k})$ and iMASP(n) can be solved exactly, their state size grows exponentially with the number of stages. To alleviate this computational difficulty, we use anytime heuristic search algorithms URTDP (Algorithm 1) and Learning Real-Time A* (Korf, 1990) to, respectively, solve $\underline{i\text{MASP}}(\frac{1}{k})$ and iMASP(n) approximately. The *adaptive greedy policy for ℓGP* repeatedly chooses a reward-maximizing action (i.e., by repeatedly solving iMASP($\frac{1}{k}$) with $t = 0$ in (5.1)) to form the observation paths. The *non-adaptive greedy policy for GP* performs likewise but does it in the log-scale. In contrast to the above policies that optimize the entropy criterion (4.1), a non-adaptive greedy policy is proposed by Guestrin *et al.* (2005) to approximately maximize the mutual information (MI) criterion for the GP; it repeatedly selects a new sampling location that maximizes the increase in MI. We call this the *MI-based policy*.

6.1 Performance Metrics

Two metrics are used to evaluate the above policies: (a) *Posterior map entropy* (ENT) $\mathbb{H}[\mathbf{Y}_{\overline{\mathbf{x}}_{0:t}}|d_t]$ of domain $\mathcal{X}$ is the criterion being optimized (4.1) that measures the posterior joint entropy of the original measurements $\mathbf{Y}_{\overline{\mathbf{x}}_{0:t}}$ at the unobserved locations $\overline{\mathbf{x}}_{0:t}$ given the posterior data d_t. For the case of 2 (1) robots, $t = 16$ (17). A smaller ENT implies lower uncertainty of the map or higher degree of information captured by the map; (b) *Mean-squared relative error* (ERR) $|\mathcal{X}|^{-1}\sum_{x\in\mathcal{X}}\{(y_x - \mu_{Y_x|d_t})/\bar{\mu}\}^2$ measures the squared relative differences between the prediction $\mu_{Y_x|d_t}$ (i.e., using the ℓGP posterior mean) and the ground truth measurement y_x averaged over all locations in $\mathcal{X}$. Although this criterion is not the one being optimized, it allows the use of ground truth measurements $\{y_x\}_{x\in\mathcal{X}}$ to evaluate if the field is being mapped accurately. A smaller ERR implies higher mapping accuracy.

6.2 Test Results

Table 6.1 shows the results of various policies with different assumed models and robot team sizes for chl-a and K fields. For $\underline{i\text{MASP}}(\frac{1}{k})$ and $i\text{MASP}(n)$, the results are obtained using the policies provided by the anytime algorithms after running 120000 simulated paths. The differences in results between policies have been verified using t-tests ($\alpha = 0.1$) to be statistically significant.

6.2.1 Plankton density data

The results show that the strictly adaptive $\underline{\pi}^{\frac{1}{k}}$ achieves lowest ENT and ERR as compared to the tested policies. From Fig. 6.1a, $\underline{\pi}^{\frac{1}{k}}$ moves the robots to sample the hotspots showing higher spatial variability whereas π^n moves them to sparsely sampled areas. Figs. 6.1b and 6.1c show, respectively, the prediction error maps resulting from $\underline{\pi}^{\frac{1}{k}}$ and π^n; the prediction error at each location x is measured using $|y_x - \mu_{Y_x|d_t}|/\bar{\mu}$. Locations with large prediction errors are mostly concentrated in the left region where the field is highly-varying and contains higher measurements. Compared to $\underline{\pi}^{\frac{1}{k}}$, π^n incurs large prediction errors at more locations in or close to hotspots, thus resulting in higher ERR.

We also compare the time needed to run the first 10000 SIMULATED-PATH(d_0, t)'s of the $\underline{i\text{MASP}}(\frac{1}{k})$-based URTDP algorithm to that of the $\underline{\text{MASP}}(\frac{1}{k})$-based URTDP algorithm, which are 115s and 10340s respectively for 2 robots (i.e., 90× faster). They, respectively, take 66s and 2835s for 1 robot (i.e., 43× faster). So, scaling to 2 robots incurs 1.73× and 3.65× more time for the respective algorithms. The induced policy $\underline{\pi}^{\frac{1}{k}}$ from solving $\underline{i\text{MASP}}(\frac{1}{k})$ can already achieve the performance reported in Table 6.1 for 2 robots, and ENT of 389.23 and ERR of 0.231 for 1 robot. In contrast, the induced policy $\underline{\pi}^{\frac{1}{k}}$ from solving $\underline{\text{MASP}}(\frac{1}{k})$ only improves to ENT of 377.82 (391.85) and ERR of 0.233 (0.252) for 2 (1) robots, which are slightly worse off.

Table 6.1: Performance comparison of information-theoretic policies for chl-a and K fields: 1R (2R) denotes 1 (2) robots.

Plankton density (chl-a) field		ENT		ERR	
Exploration policy	Model	1R	2R	1R	2R
Adaptive $\underline{\pi}^{1/k}$	ℓGP	381.37	376.19	0.1827	0.2319
Adaptive greedy	ℓGP	382.97	383.55	0.2919	0.2579
Non-adaptive π^n	GP	390.62	399.63	0.4145	0.3194
Non-adaptive greedy	GP	392.35	392.51	0.2994	0.3356
MI-based	GP	395.37	397.02	0.2764	0.2706
Potassium (K) field		ENT		ERR	
Exploration policy	Model	1R	2R	1R	2R
Adaptive $\underline{\pi}^{1/k}$	ℓGP	47.330	48.287	0.0299	0.0213
Adaptive greedy	ℓGP	61.080	56.181	0.0457	0.0302
Non-adaptive π^n	GP	67.084	59.318	0.0434	0.0358
Non-adaptive greedy	GP	58.704	64.186	0.0431	0.0335
MI-based	GP	59.058	67.390	0.0435	0.0343

6.2.2 Potassium distribution data

The results show again that $\underline{\pi}^{\frac{1}{k}}$ achieves lowest ENT and ERR. From Fig. 6.1d, $\underline{\pi}^{\frac{1}{k}}$ again moves the robots to sample the hotspots showing higher spatial variability whereas π^n moves them to sparsely sampled areas. Compared to $\underline{\pi}^{\frac{1}{k}}$, π^n incurs large prediction errors at a greater number of locations in or close to hotspots as shown in Figs. 6.1e and 6.1f, thus resulting in higher ERR.

To run 10000 SIMULATED-PATH(d_0, t)'s, the $\underline{i\text{MASP}}(\frac{1}{k})$-based URTDP algorithm is 84× (48×) faster than that of the $\underline{\text{MASP}}(\frac{1}{k})$-based URTDP algorithm for 2 (1) robots. Scaling to 2 robots incurs 1.93× and 3.37× more time for the respective algorithms. The induced policy $\underline{\pi}^{\frac{1}{k}}$ from solving $\underline{i\text{MASP}}(\frac{1}{k})$ can already achieve the performance reported in Table 6.1 for 1 and 2 robots. In contrast, the induced policy $\underline{\pi}^{\frac{1}{k}}$ from solving $\underline{\text{MASP}}(\frac{1}{k})$ achieves worse ENT of 67.132 (55.015) for 2 (1) robots. It achieves worse ERR of 0.032 for 2 robots but better ERR of 0.025 for 1 robot.

6.2.3 Summary of test results

The above results show that the strictly adaptive $\underline{\pi}^{\frac{1}{k}}$ can learn the highest-quality hotspot field map (i.e., lowest ENT and ERR) among the tested state-of-the-art strategies. After evaluating whether MASP- vs. iMASP-based anytime planners are time-efficient for real-time deployment, we observe that the induced policy $\underline{\pi}^{\frac{1}{k}}$ from solving $\underline{i\text{MASP}}(\frac{1}{k})$ can achieve mapping performance comparable to the induced policy $\underline{\pi}^{\frac{1}{k}}$ from solving $\underline{\text{MASP}}(\frac{1}{k})$ using significantly less time, and the incurred planning time is also less sensitive to larger robot team size. Lastly, we see in Fig. 6.1 that the strictly adaptive $\underline{\pi}^{\frac{1}{k}}$ has exploited clustering phenomena (i.e., hotspots) to achieve lower ENT and ERR than that of the non-adaptive π^n.

Chapter 7

Quantifying "Hotspotness"

It is of practical interest to be able to quantitatively characterize the "hotspotness" of an environmental field. In this manner, environmental fields of varying degrees of "hotspotness" can be prescribed accordingly, that is, by assigning high degrees of "hotspotness" to fields with pronounced hotspots and low degrees of "hotspotness" to smoothly-varying fields. This "hotspotness" measure can then be used, for example, in environmental sensing applications (e.g., monitoring of algal bloom or pollution) to (a) indicate the severity and extent of contamination and the subsequent remediation action or (b) rank a list of potential regions sampled at low resolution (e.g., via remote sensing) and select the region with the highest degree of "hotspotness" to perform *in situ* high-resolution adaptive sampling.

In this chapter, we propose a novel "hotspotness" measure, which is defined in terms of the spatial correlation properties of the hotspot field (Section 7.1). Specifically, by assuming the hotspot field to vary as a realization of the ℓGP (Sections 3.5.2 and 4.3.2), its spatial correlation properties can be represented by the hyperparameters of the ℓGP covariance structure. This then allows the proposed "hotspotness" index to be defined using the hyperparameters. Through the use of the hyperparameters, we will discuss how the "hotspotness" index can be related to the intensity, size, and diffuseness of the hotspots in the environmental field (Section 7.2). In Section 7.3, we apply the "hotspotness" index to

a real-world phosphorus distribution field. The "hotspotness" index is generalized in Section 7.4 to include a parameter that can be calibrated to penalize a certain class of fields with "subjectively" higher than expected degrees of "hotspotness".

7.1 Index of "Hotspotness"

Traditionally, a hotspot is defined as a location in which its corresponding measurement exceeds a pre-defined threshold (De Oliveira and Ecker, 2002; Long and Wilson, 1997). This threshold is expected to be considerably higher than the mean of the field measurements. However, hotspot locations do not usually occur in isolation; the neighboring locations of a hotspot are also likely to be hotspots. So, it will be useful to, instead, consider a hotspot as a cluster of spatially connected locations with measurements exceeding the pre-defined threshold and consequently characterize a hotspot field with spatial correlation properties.

As assumed in Sections 3.5.2 and 4.3.2, we define a hotspot field to vary as a realization of the ℓGP. Then, the spatial correlation properties of a hotspot field can be represented by that of the ℓGP, in particular, its covariance structure. Specifically, we define the covariance structure using the squared exponential covariance function

$$\sigma_{Z_x Z_u} \stackrel{\triangle}{=} \sigma_s^2 \exp\left\{-\frac{||x-u||^2}{2\ell^2}\right\} + \sigma_n^2 \delta_{xu} \tag{7.1}$$

for $x, u \in \mathcal{X}$ with the signal variance σ_s^2, the noise variance σ_n^2, and the length-scale ℓ acting as hyperparameters, and δ_{xu} is a Kronecker delta of value 1 if $x = u$, and 0 otherwise. Since the covariance function is isotropic, $\sigma_{Z_x Z_u}$ is a function of $r \stackrel{\triangle}{=} ||x-u||$ and can therefore be written as a function of just a single argument, i.e., $\sigma(r)$:

$$\sigma(r) \stackrel{\triangle}{=} \sigma_s^2 \exp\left\{-\frac{r^2}{2\ell^2}\right\} + \sigma_n^2 \delta_r \tag{7.2}$$

where δ_r is a Kronecker delta of value 1 if $r = 0$, and 0 otherwise.

Definition 7.1.1 ("Hotspotness" Index). The degree of "hotspotness", denoted by H_θ, of an environmental field is defined as

$$\begin{aligned} H_\theta &\triangleq \int_0^{\ell\sqrt{2\ln\frac{1}{\rho\theta}}} \frac{\sigma(r)}{\sigma_s^2+\sigma_n^2} - \theta \; dr \\ &= \ell\left\{\frac{\sqrt{0.5\pi}}{\rho}\,\mathrm{erf}\left(\sqrt{\ln\frac{1}{\rho\theta}}\right) - \theta\sqrt{2\ln\frac{1}{\rho\theta}}\right\} \end{aligned}$$

where $\rho = 1 + \dfrac{\sigma_n^2}{\sigma_s^2}$ and $0 \le \theta \le \dfrac{1}{\rho}$.

Remarks.

1. H_θ measures the area bounded below the normalized squared exponential curve $\dfrac{\sigma(r)}{\sigma_s^2+\sigma_n^2}$, above the horizontal line of value θ, and to the right of the vertical axis as depicted in Fig. 7.1;
2. $H_\theta \ge 0$;
3. $H_0 \propto \dfrac{\ell}{\rho}$; and
4. $\theta_1 > \theta_2 \Leftrightarrow H_{\theta_1} < H_{\theta_2}$.

For the case of $\theta = 0$, we can observe from Remark 3 of Definition 7.1.1 that a high degree of "hotspotness" H_0 is achieved with a large length-scale ℓ and a small ρ (i.e., small noise-to-signal ratio σ_n^2/σ_s^2).

Using the "hotspotness" index H_0, it is possible for a hotspot field with both length-scale ℓ and noise-to-signal ratio σ_n^2/σ_s^2 larger than that of another field to achieve the same degree of "hotspotness". For example, a slowly-varying field with high degree of noise is described

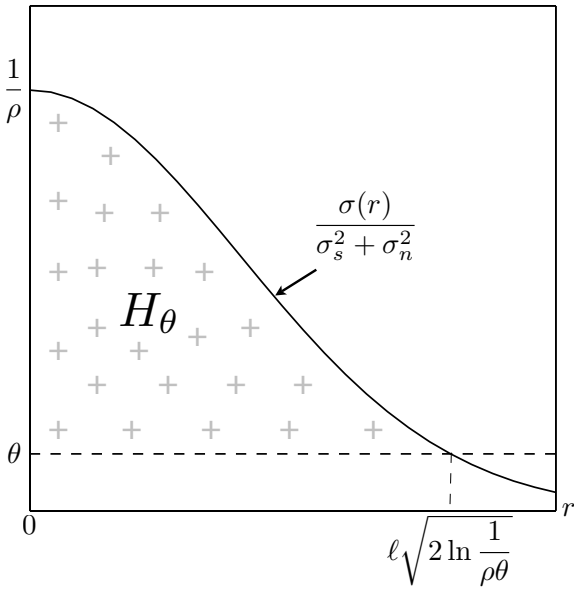
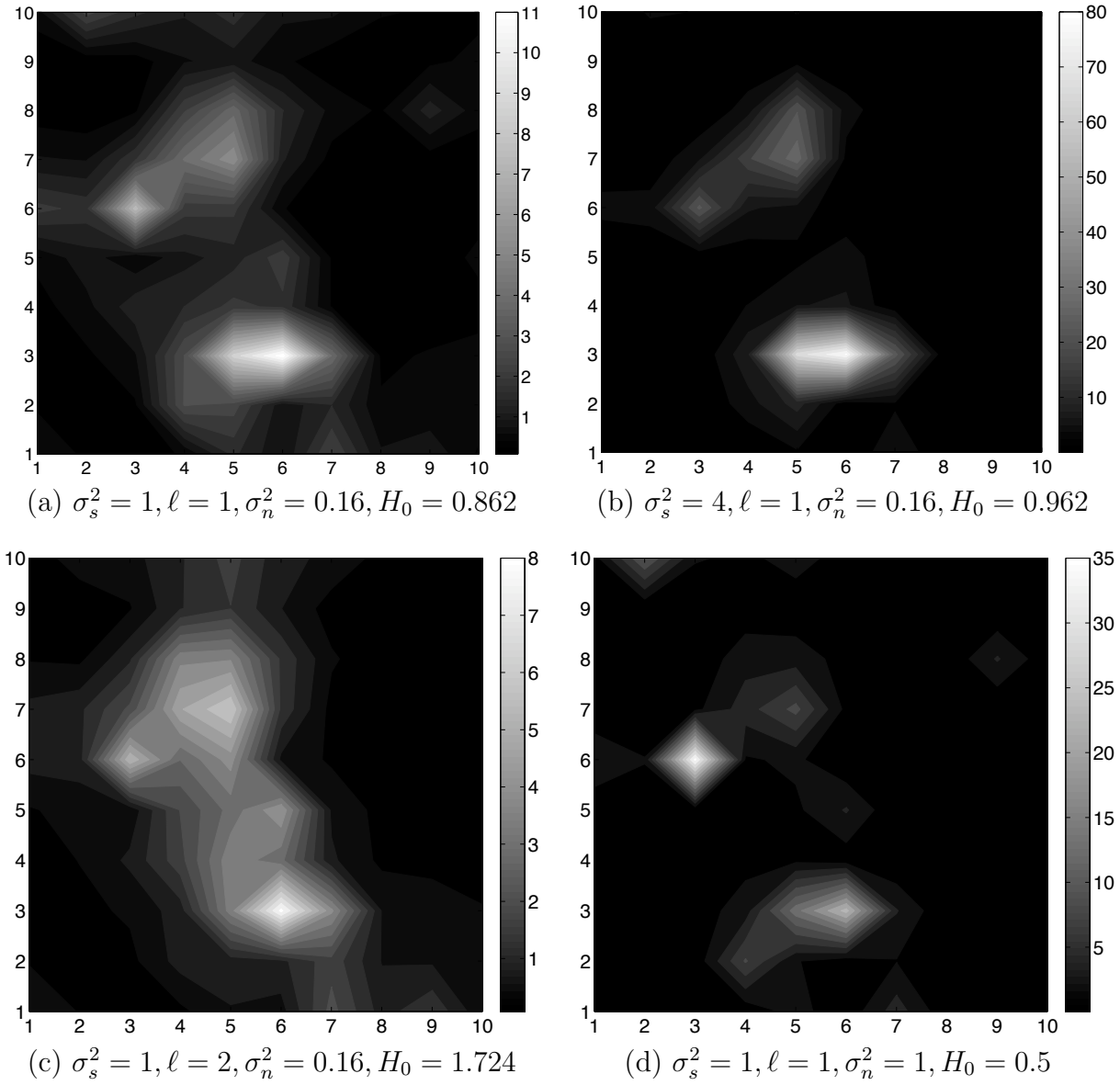


Figure 7.1: Graphical interpretation of the degree of "hotspotness" H_θ measuring the bounded area covered with '+'s. Refer to definition 7.1.1 for more details.



(a) $\sigma_s^2 = 1, \ell = 1, \sigma_n^2 = 0.16, H_0 = 0.862$

(b) $\sigma_s^2 = 4, \ell = 1, \sigma_n^2 = 0.16, H_0 = 0.962$

(c) $\sigma_s^2 = 1, \ell = 2, \sigma_n^2 = 0.16, H_0 = 1.724$

(d) $\sigma_s^2 = 1, \ell = 1, \sigma_n^2 = 1, H_0 = 0.5$

Figure 7.2: Hotspot field simulations via ℓGP with varying hyperparameters producing different characteristics of hotspots and degrees of "hotspotness" H_0. See text for explanation.

by very large length-scale ℓ and noise-to-signal ratio σ_n^2/σ_s^2 while a rapidly-varying field with low degree of noise is characterized by very small length-scale ℓ and noise-to-signal ratio σ_n^2/σ_s^2, both of which may produce similarly low degrees of "hotspotness".

7.2 Effects of Spatial Correlation on Hotspot Characteristics and "Hotspotness"

Fig. 7.2 illustrates how the hyperparameters (i.e., length-scale, signal variance, and noise variance) can be varied to produce different characteristics of hotspots and degrees of "hotspotness" H_0. Fig. 7.2a shows a simulated ℓGP field with unperturbed hyperparameters. Increasing the signal variance σ_s^2 amplifies the intensity of hotspots (Fig. 7.2b), which consequently become more pronounced. Raising the length-scale ℓ increases the size of hotspots (Fig. 7.2c), thus widening them. Lastly, increasing the noise variance σ_n^2 makes the hotspots more diffuse (Fig. 7.2d). Notice from Figs. 7.2b and 7.2c that raising the signal variance σ_s^2 or length-scale ℓ increases the degree of "hotspotness" H_0 while raising the noise variance σ_n^2 decreases the degree of "hotspotness" H_0 as shown in Fig. 7.2d. Therefore, when the hotspots in the environmental field are intense/pronounced, wide, and undiffused, a high degree of "hotspotness" H_0 is expected.

7.3 Application: Phosphorus Distribution Field

In Fig. 7.3, we demonstrate the application of the "hotspotness" index H_0 to a real-world dataset (i.e., the phosphorus distribution field of Broom's Barn farm) and show the degrees of "hotspotness" of the field in different sub-regions. It can be observed that the area with the highest degree of "hotspotness" (i.e., $H_0 = 1.7142$) features a pronounced hotspot while the area with the lowest degree of "hotspotness" (i.e., $H_0 = 0.9542$) exhibits a smoothly-varying

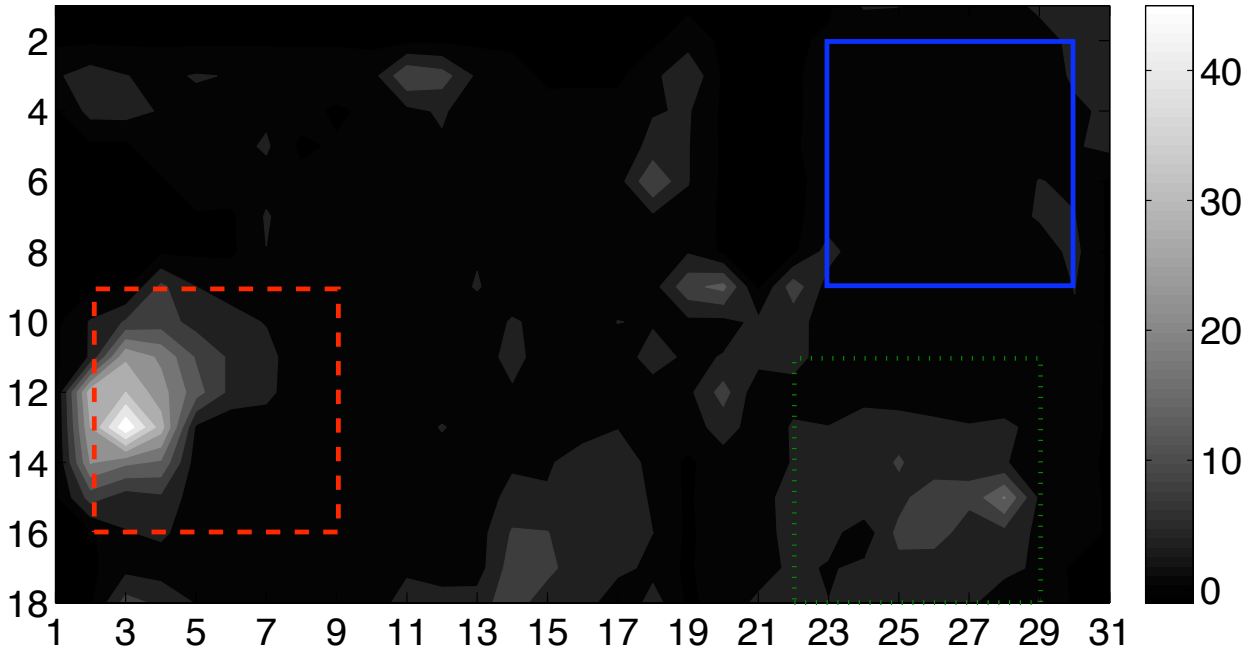


Figure 7.3: Phosphorus distribution field with $H_0 = 1.7142$, $H_0 = 1.2486$, and $H_0 = 0.9542$ for the sub-regions boxed with a dashed red line, a dotted green line, and a solid blue line respectively.

field.

7.4 Generalized Index of "Hotspotness"

When comparing the degrees of "hotspotness" of different fields using the index H_0, we may perceive a field with larger length-scale ℓ and noise-to-signal ratio σ_n^2/σ_s^2 to be producing "subjectively" higher than expected degree of "hotspotness" H_0. Such a field can, for example, be a slowly-varying field with exceedingly large length-scale but not as large noise-to-signal ratio, thus resulting in a relatively high degree of "hotspotness" H_0 that may be undesirable.

To remedy this, we propose a generalized "hotspotness" measure H_θ. It is calibrated by

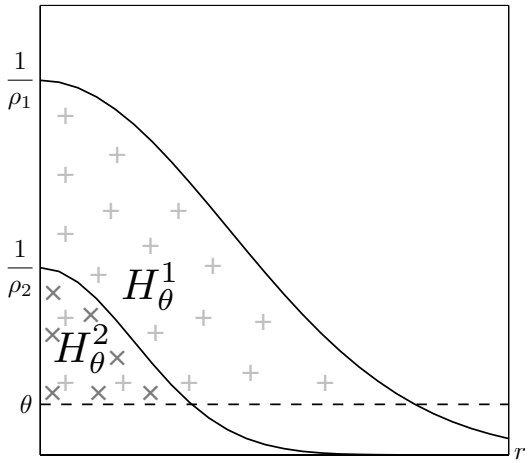
setting $\theta > 0$, which controls the reduction of the degree of "hotspotness" of all fields in comparison. This can be observed in Fig. 7.1 such that increasing θ reduces the degree of "hotspotness" H_θ. More importantly, a field with larger length-scale ℓ and noise-to-signal ratio σ_n^2/σ_s^2 is penalized more because its degree of "hotspotness" is reduced by a greater extent as shown in Fig. 7.4c. As a result, its "subjectively" higher than expected degree of "hotspotness" H_0 can be reduced to be lower than that of another field with smaller length-scale ℓ and noise-to-signal ratio σ_n^2/σ_s^2 after calibration. This is guaranteed by the following result, in particular, that arising from condition 4:

Theorem 7.4.1. Let the degree of "hotspotness" of field i be H_θ^i with hyperparameters ℓ_i and ρ_i. For $\theta > 0$, $H_\theta^1 > H_\theta^2$ if one of the following conditions is satisfied:

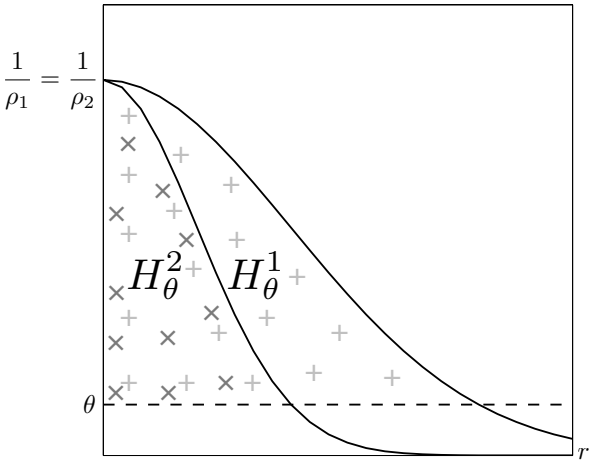
1. $\ell_1 \geq \ell_2$ and $\rho_1 < \rho_2$;
2. $\ell_1 > \ell_2$ and $\rho_1 = \rho_2$;
3. $H_0^1 \geq H_0^2$ and $\ell_1 < \ell_2$ and $\rho_1 < \rho_2$;
4. $H_0^1 < H_0^2$ and $\ell_1 < \ell_2$ and $\rho_1 < \rho_2$ and $\theta \geq \dfrac{1}{\rho_1}\left(\dfrac{\rho_1}{\rho_2}\right)^{\frac{1}{1-(\ell_1/\ell_2)^2}}$.

To prove the above result, observe from Fig. 7.4 that the four conditions give rise to different pairwise sets of hotspot fields with differing hyperparameters and degrees of "hotspotness" H_0^1 and H_0^2, which can achieve $H_\theta^1 > H_\theta^2$ after calibration. Conditions 1 and 2 are depicted respectively in Figs. 7.4a and 7.4b, from which we can clearly see that $H_\theta^1 > H_\theta^2$ for $\theta \geq 0$. Condition 3 is illustrated in Fig. 7.4c; when $H_0^1 \geq H_0^2$, increasing θ reduces the degree of "hotspotness" of field 2 more than that of field 1. Hence, $H_\theta^1 > H_\theta^2$ for $\theta \geq 0$. The most interesting result would be that arising from condition 4: from Fig. 7.4c, we can

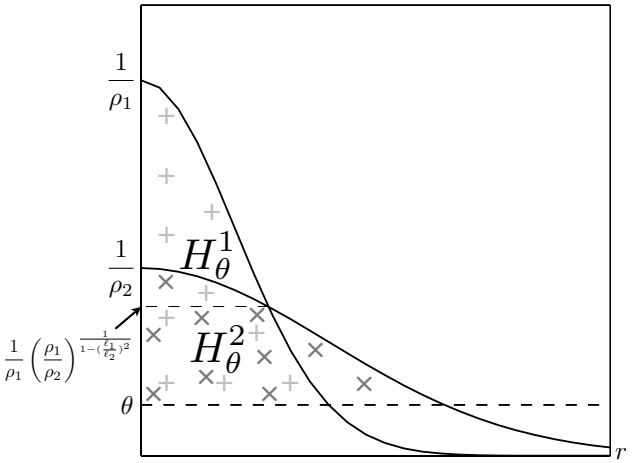
see that field 2 with length-scale ℓ_2, noise-to-signal ratio or ρ_2, and degree of "hotspotness" H_0^2 larger than that of field 1 can achieve lower degree of "hotspotness" H_θ^2 after calibrating with a large enough θ. That is, $\theta \geq \frac{1}{\rho_1}\left(\frac{\rho_1}{\rho_2}\right)^{\frac{1}{1-(\ell_1/\ell_2)^2}}$ is a sufficient condition for $H_\theta^1 > H_\theta^2$.



(a) $\ell_1 \geq \ell_2, \rho_1 < \rho_2$



(b) $\ell_1 > \ell_2, \rho_1 = \rho_2$



(c) $\ell_1 < \ell_2, \rho_1 < \rho_2$

Figure 7.4: Graphical interpretations of H_θ^1 and H_θ^2 measuring the bounded areas covered, respectively, by '+'s and '×'s. See the proof of Theorem 7.4.1 for more details.

Chapter 8

Effects of Spatial Correlation on Performance Advantage of Adaptivity

In Section 3.3, we have shown that the optimal adaptive policy π^1 from solving MASP(1) (iMASP(1)) performs better than or at least as well as the optimal non-adaptive policy π^n from solving MASP(n) (iMASP(n)) . But, this result does not quantify the extent of such a performance advantage. In this chapter, we will investigate how the spatial correlation properties of the hotspot field (in particular, the length-scale hyperparameter ℓ of the ℓGP covariance structure) affect the extent of this advantage. We assume here that the ℓGP covariance structure is defined by the squared exponential covariance function (7.2).

We first show in Section 8.1 that for white-noise process fields (i.e., $\ell = 0$) or constant fields (i.e., $\ell = \infty$), the multi-stage adaptive MASP(1) and iMASP(1) provide no performance advantage over the non-adaptive MASP(n) and iMASP(n), respectively. Then, we show in Section 8.2 that the performance advantage of the 2-stage adaptive iMASP(1) is zero if and only if the field is constant or a white-noise process. Note that the contrapositive of this second result implies the performance advantage is positive if and only if the length-scale is non-extreme. Lastly, we illustrate that the performance advantage of the 2-stage adaptive iMASP(1) improves with decreasing noise-to-signal variance ratio and peaks

at some intermediate length-scale.

8.1 Multi-Stage MASP(1) and iMASP(1)

The first result indicates that sampling white-noise process fields (i.e., $\ell = 0$) or constant fields (i.e., $\ell = \infty$) are sufficient conditions for multi-stage adaptive exploration to yield no performance advantage under the mean-squared error (3.2) or entropy (4.1) criterion:

Theorem 8.1.1. If $\ell = 0$ or $\ell = \infty$, MASP(1) and iMASP(1) can be reduced to be single-staged, and their optimal adaptive policies can be reduced to be non-adaptive.

The proof of the above result is in Appendix A.14. To understand the intuition behind Theorem 8.1.1, prior observations made in a white-noise process field offer no information on the unobserved locations. So, the policy π^1 to select new unobserved locations becomes independent of previous observations and can thus be reduced to be non-adaptive. In a constant field, any new observation considered during a stagewise selection will provide the same amount of information on the unobserved locations. As a result, the policy π^1 can also be reduced to be non-adaptive.

8.2 2-Stage iMASP(1)

We know from Section 8.1 that adaptivity provides no performance advantage for extreme length-scales. Does adaptivity then offer positive performance advantage for non-extreme length-scales? Theorem 8.1.1 ensues by exploiting the simplified covariance structure due to extreme length-scales (Appendix A.14). However, the covariance structure cannot, in general, be simplified with a non-extreme length-scale. In this section, we will use a different

approach to analyze how the spatial correlation properties of the hotspot field (i.e., the length-scale, signal variance, and noise variance hyperparameters[1] of the ℓGP covariance structure) affect the performance advantage. In particular, we will evaluate the performance advantage of the 2-stage strictly adaptive iMASP(1) over the non-adaptive iMASP(n) (i.e., $n = 2$) for the case of $k = 1$ robot.

8.2.1 Exact closed-form solution for adaptive iMASP(1)

When $n = 2$, the reward-maximizing iMASP(1) (4.6) comprises the following 2-stage dynamic programming equations:

$$\begin{aligned} U_0^{\pi^1}(d_0) &= \max_{a_0 \in \mathcal{A}(x_0)} \mathbb{H}[Y_{x_1} \mid d_0] + \int f(z_{x_1} \mid d_0)\; U_1^{\pi^1}(d_1)\; \mathrm{d}z_{x_1} \\ &= \max_{a_0 \in \mathcal{A}(x_0)} \frac{1}{2} \log 2\pi e \sigma^2_{Z_{x_1}|d_0} + \mu_{Z_{x_1}|d_0} + \int f(z_{x_1} \mid d_0)\; U_1^{\pi^1}(d_1)\; \mathrm{d}z_{x_1} \\ U_1^{\pi^1}(d_1) &= \max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{x_2} \mid d_1] \\ &= \max_{a_1 \in \mathcal{A}(x_1)} \frac{1}{2} \log 2\pi e \sigma^2_{Z_{x_2}|d_1} + \mu_{Z_{x_2}|d_1} \end{aligned} \tag{8.1}$$

where $x_{i+1} = \tau(x_i, a_i)$ and $Z_{x_{i+1}} = \log Y_{x_{i+1}}$ for $i = 0, 1$. It is computationally feasible to solve (8.1) if the integral can be evaluated in closed form: observe that there are $|\mathcal{A}(x_1)|$ possible reward functions $\mathbb{H}[Y_{x_2} \mid d_1] = \mathbb{H}[Y_{x_2} \mid d_0, x_1, z_{x_1}]$ (one for each $a_1 \in \mathcal{A}(x_1)$), each of which is a linear function of z_{x_1} as shown in Appendix A.15. Therefore, a different value of z_{x_1} can induce a different reward function $\mathbb{H}[Y_{x_2} \mid d_1]$ to dominate due to the maximum operator. More specifically, $U_1^{\pi^1}(d_1)$ is a piecewise-linear function of z_{x_1} that is constituted by at most $|\mathcal{A}(x_1)|$ reward functions $\mathbb{H}[Y_{x_2} \mid d_1]$ dominating different disjoint intervals of z_{x_1}. These dominating reward functions intersect at *breakpoints*, which form the interval endpoints and can be determined exactly.

By identifying these breakpoints, the integral can be evaluated in closed form: let

[1] See (7.2) and Section 7.2 for explanation of these hyperparameters of the ℓGP covariance structure.

the support of Z_{x_1} given the sampled data d_0 be partitioned by previously established breakpoints into ν (i.e., $\nu \leq |\mathcal{A}(x_1)|$) disjoint, consecutive intervals $\mathcal{Z}_{x_1}^{[1]}, \ldots, \mathcal{Z}_{x_1}^{[\nu]}$. Suppose that each disjoint interval $\mathcal{Z}_{x_1}^{[i]}$ is associated with the dominating reward function $\mathbb{H}[Y_{x_2^{[i]}} \mid d_1] \stackrel{\triangle}{=} \max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{x_2} \mid d_1] \geq \mathbb{H}[Y_{x_2} \mid d_1]$ for any $z_{x_1} \in \mathcal{Z}_{x_1}^{[i]}$. That is, $x_2^{[i]} = \tau\left(x_1, \arg\max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{\tau(x_1,a_1)} \mid d_1]\right)$ for any $z_{x_1} \in \mathcal{Z}_{x_1}^{[i]}$. Then, the integral can be expressed in terms of the dominating reward functions for these intervals:

$$\begin{aligned}
\int f(z_{x_1} \mid d_0)\; U_1^{\pi^1}(d_0, x_1, z_{x_1})\; \mathrm{d}z_{x_1} &= \int f(z_{x_1} \mid d_0) \max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{x_2} \mid d_0, x_1, z_{x_1}]\; \mathrm{d}z_{x_1} \\
&= \sum_{i=1}^{\nu} \int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0) \max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{x_2} \mid d_0, x_1, z_{x_1}]\; dz_{x_1} \\
&= \sum_{i=1}^{\nu} \int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0)\; \mathbb{H}[Y_{x_2^{[i]}} \mid d_0, x_1, z_{x_1}]\; dz_{x_1}\;,
\end{aligned} \tag{8.2}$$

which can be reduced to a closed-form expression as shown in Appendix A.15. Consequently, the exact closed-form solution of $U_0^{\pi^1}(d_0)$ can be obtained.

8.2.2 Exact closed-form solution for non-adaptive iMASP(n)

On the other hand, the reward-maximizing iMASP(n) (4.5) with $n = 2$ has the following form:

$$\begin{aligned}
U_0^{\pi^2}(d_0) &= \max_{\mathbf{a}_{0:1}} \mathbb{H}[\mathbf{Y}_{\mathbf{x}_{1:2}} \mid d_0] \\
&= \max_{a_0 \in \mathcal{A}(x_0)} \mathbb{H}[Y_{x_1} \mid d_0] + \max_{a_1 \in \mathcal{A}(x_1)} \int f(z_{x_1} \mid d_0)\; \mathbb{H}[Y_{x_2} \mid d_1]\; \mathrm{d}z_{x_1} \\
&= \max_{a_0 \in \mathcal{A}(x_0)} \mathbb{H}[Y_{x_1} \mid d_0] + \int f(z_{x_1} \mid d_0)\; \mathbb{H}[Y_{x_2^*} \mid d_1]\; \mathrm{d}z_{x_1}
\end{aligned} \tag{8.3}$$

where $\mathbf{x}_{1:2} = \tau(\mathbf{x}_{0:1}, \mathbf{a}_{0:1})$ and $x_2^* = \tau\left(x_1, \arg\max_{a_1 \in \mathcal{A}(x_1)} \int f(z_{x_1} \mid d_0)\; \mathbb{H}[Y_{\tau(x_1,a_1)} \mid d_1]\; \mathrm{d}z_{x_1}\right)$ for a given location x_1.

8.2.3 Performance advantage of adaptive exploration

To evaluate the performance advantage of adaptivity $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0)$, we need the following lemma:

Lemma 8.2.1. Let $(w_0(x), w_1(x)) \triangleq \Sigma_{x\mathbf{x}_{0:1}}\Sigma^{-1}_{\mathbf{x}_{0:1}\mathbf{x}_{0:1}}$. Then,

$$w_0(x) = \frac{\sigma_{Z_{x_0}Z_x} - w_1(x)\sigma_{Z_{x_0}Z_{x_1}}}{\sigma^2_{Z_{x_0}}} \quad \text{and} \quad w_1(x) = \frac{\sigma_{Z_{x_1}Z_x|d_0}}{\sigma^2_{Z_{x_1}|d_0}} \ .$$

The above lemma provides an alternative interpretation of the weights on the observed measurements $\mathbf{z}_{\mathbf{x}_{0:1}}$ when they are used to predict the measurement z_x of an unobserved location x via the Gaussian posterior mean (3.19). In particular, the original weight vector $\Sigma_{x\mathbf{x}_{0:1}}\Sigma^{-1}_{\mathbf{x}_{0:1}\mathbf{x}_{0:1}}$ is decomposed into individual weights that are expressed in terms of Gaussian posterior covariances and variances. As a result, the individual weights can be interpreted more easily as opposed to understanding them directly through the inverse covariance matrix $\Sigma^{-1}_{\mathbf{x}_{0:1}\mathbf{x}_{0:1}}$. This lemma can be generalized to cater to the weight vector $\Sigma_{x\mathbf{x}_{0:n}}\Sigma^{-1}_{\mathbf{x}_{0:n}\mathbf{x}_{0:n}}$.

Before we can determine the performance advantage of adaptivity $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0)$, we have to first consider its performance advantage when the same location x_1 is already selected to be explored by both adaptive iMASP(1) and non-adaptive iMASP(n), and then work backwards. Supposing location x_1 is selected to be explored, it can be noted from (8.1) and (8.2) that the strictly adaptive iMASP(1) waits and observes the corresponding measurement z_{x_1} before selecting the next location x_2 to explore. On the other hand, we can see from (8.3) that the non-adaptive iMASP(n) selects the next location x_2 before observing z_{x_1}. Given the same selected location x_1 to be explored by both adaptive iMASP(1) and non-adaptive iMASP(n), the resulting performance advantage of adaptivity, denoted by

$D(x_1, d_0)$, is then

$$
\begin{aligned}
& D(x_1, d_0) \\
& \stackrel{\triangle}{=} \mathbb{H}[Y_{x_1} \mid d_0] + \int f(z_{x_1} \mid d_0) \max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{x_2} \mid d_1] \, \mathrm{d}z_{x_1} - \\
& \quad \left(\mathbb{H}[Y_{x_1} \mid d_0] + \max_{a_1 \in \mathcal{A}(x_1)} \int f(z_{x_1} \mid d_0) \, \mathbb{H}[Y_{x_2} \mid d_1] \, \mathrm{d}z_{x_1} \right) \\
& = \int f(z_{x_1} \mid d_0) \max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{x_2} \mid d_1] \, \mathrm{d}z_{x_1} - \max_{a_1 \in \mathcal{A}(x_1)} \int f(z_{x_1} \mid d_0) \, \mathbb{H}[Y_{x_2} \mid d_1] \, \mathrm{d}z_{x_1} \\
& = \sum_{i=1}^{\nu} \int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0) \max_{a_1 \in \mathcal{A}(x_1)} \mathbb{H}[Y_{x_2} \mid d_0, x_1, z_{x_1}] \, dz_{x_1} - \int f(z_{x_1} \mid d_0) \, \mathbb{H}[Y_{x_2^*} \mid d_1] \, \mathrm{d}z_{x_1} \\
& = \sum_{i=1}^{\nu} \int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0) \, \mathbb{H}[Y_{x_2^{[i]}} \mid d_1] \, dz_{x_1} - \sum_{i=1}^{\nu} \int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0) \, \mathbb{H}[Y_{x_2^*} \mid d_1] \, dz_{x_1} \\
& = \sum_{i=1}^{\nu} \int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0) \, \mathbb{H}[Y_{x_2^{[i]}} \mid d_1] - \mathbb{H}[Y_{x_2^*} \mid d_1] \, dz_{x_1} \\
& = \sum_{i=1}^{\nu} \int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0) \, \left(w_1(x_2^{[i]}) z_{x_1} + c(d_0, x_1, x_2^{[i]}) \right) - \left(w_1(x_2^*) z_{x_1} + c(d_0, x_1, x_2^*) \right) dz_{x_1} \\
& = \sum_{i=1}^{\nu} \underline{p}_{x_1}^{[i]} \left[\left(w_1(x_2^{[i]}) - w_1(x_2^*) \right) \underline{z}_{x_1}^{[i]} + c(d_0, x_1, x_2^{[i]}) - c(d_0, x_1, x_2^*) \right] \\
& = \sum_{i=1}^{\nu} \underline{p}_{x_1}^{[i]} \left(\mathbb{H}[Y_{x_2^{[i]}} \mid d_0, x_1, \underline{z}_{x_1}^{[i]}] - \mathbb{H}[Y_{x_2^*} \mid d_0, x_1, \underline{z}_{x_1}^{[i]}] \right)
\end{aligned}
\tag{8.4}
$$

where $c(d_0, x_1, x_2) = \frac{1}{2} \log 2\pi e \sigma^2_{Z_{x_2}|d_1} + (1 - w_1(x_2)) \mu_{Z_{x_2}} + w_0(x_2)(z_{x_0} - \mu)$, and $\underline{p}_{x_1}^{[i]}$ and $\underline{z}_{x_1}^{[i]}$ are defined according to that of $\underline{i\mathrm{MASP}}(\frac{1}{k})$ (5.4). The second to fourth equalities are due to (8.2). From the expression under the fifth equality in (8.4), we know that for any i, $\mathbb{H}[Y_{x_2^{[i]}} \mid d_1] \geq \mathbb{H}[Y_{x_2^*} \mid d_1]$ for any $z_{x_1} \in \mathcal{Z}_{x_1}^{[i]}$, thus resulting in $D(x_1, d_0) \geq 0$ for any x_1. The sixth and last equalities follow from

$$
\begin{aligned}
\mathbb{H}[Y_{x_2} \mid d_1] = & \frac{1}{2} \log 2\pi e \sigma^2_{Z_{x_2}|d_1} + \mu_{Z_{x_2}|d_1} \\
= & \frac{1}{2} \log 2\pi e \sigma^2_{Z_{x_2}|d_1} + (1 - w_0(x_2) - w_1(x_2)) \mu_{Z_{x_2}} + w_0(x_2) z_{x_0} + w_1(x_2) z_{x_1} \\
= & \; c(d_0, x_1, x_2) + w_1(x_2) z_{x_1}
\end{aligned}
\tag{8.5}
$$

such that the second equality in (8.5) is due to Lemma 8.2.1. It can be observed that $\mathbb{H}[Y_{x_2} \mid d_1]$ is a linear function of z_{x_1}. The seventh equality in (8.4) is due to linearity of expectation.

Now, we remove the assumption of having to select the same location x_1 to be explored and evaluate the performance advantage $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0)$ of the strictly adaptive iMASP(1) over the non-adaptive iMASP(n). We know that this performance advantage is loosely bounded by

$$\min_{a_0 \in \mathcal{A}(x_0)} D(\tau(x_0, a_0), d_0) \leq U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0) \leq \max_{a_0 \in \mathcal{A}(x_0)} D(\tau(x_0, a_0), d_0) \ . \tag{8.6}$$

The next lemma follows directly from (8.6):

Lemma 8.2.2. If $D(\tau(x_0, a_0), d_0) = 0$ for every $a_0 \in \mathcal{A}(x_0)$, $U_0^{\pi^1}(d_0) = U_0^{\pi^2}(d_0)$.

The following lemma provides a sufficient condition for satisfying the antecedent of Lemma 8.2.2:

Lemma 8.2.3. Given x_1, if $w_1(\tau(x_1, a_1)) = w_1(\tau(x_1, \tilde{a}_1))$ for every pair of actions $a_1, \tilde{a}_1 \in \mathcal{A}(x_1)$, $D(x_1, d_0) = 0$.

To prove the above result, we know from (8.5) that $w_1(\tau(x_1, a_1))$ is the gradient of the line $\mathbb{H}[Y_{x_2} \mid d_1] = \mathbb{H}[Y_{\tau(x_1,a_1)} \mid d_0, x_1, z_{x_1}]$. So, the antecedent of Lemma 8.2.3 implies that the lines $\mathbb{H}[Y_{\tau(x_1,a_1)} \mid d_0, x_1, z_{x_1}]$ for $a_1 \in \mathcal{A}(x_1)$ are all parallel. As a result, the line $\mathbb{H}[Y_{x_2^*} \mid d_1]$ has to dominate the entire support of Z_{x_1}, which implies $D(x_1, d_0) = 0$ by (8.4).

The next lemma tells us that an extreme length-scale is a sufficient condition for satisfying the antecedent of Lemma 8.2.3:

Lemma 8.2.4. Given x_1, if $\ell = 0$ or $\ell = \infty$, $w_1(\tau(x_1, a_1)) = w_1(\tau(x_1, \tilde{a}_1))$ for every pair of actions $a_1, \tilde{a}_1 \in \mathcal{A}(x_1)$.

To prove the above lemma, recall from Lemma 8.2.1 that $w_1(x_2) = \dfrac{\sigma_{Z_{x_1} Z_{x_2}|d_0}}{\sigma^2_{Z_{x_1}|d_0}}$. If $\ell = 0$, (A.13) can be used to show that $w_1(x_2) = 0$ for every x_2. If $\ell = \infty$, (A.14) can be used to show that $w_1(x_2)$ evaluates to the same constant for every x_2. Hence, the lemma follows.

The next result follows immediately from Lemmas 8.2.2, 8.2.3, and 8.2.4. It indicates that an extreme length-scale is a sufficient condition for adaptivity to offer zero performance advantage:

Theorem 8.2.1. If $\ell = 0$ or $\ell = \infty$, $U_0^{\pi^1}(d_0) = U_0^{\pi^2}(d_0)$.

We note that the same outcome is reached in Theorem 8.1.1, but that result holds for the more general multi-stage case. The contrapositive of Theorem 8.2.1 indicates that a non-extreme length-scale is a necessary condition for adaptivity to yield positive performance advantage.

We will now demonstrate how the performance advantage of adaptivity can be positive (i.e., $U_0^{\pi^1}(d_0) > U_0^{\pi^2}(d_0)$) for non-extreme length-scales. Recall from Section 3.1 that the hotspot field is discretized into a grid of sampling cell locations (e.g., Fig. 6.1) such that each cell's width is assumed to be 1 m^2. We assume that the robot's actions are restricted to moving to the front, left, or right cell: when the robot moves either forward, left, or right from its current cell location x_i, its new cell location is denoted by x_{i+1}^f, x_{i+1}^ℓ, and x_{i+1}^r respectively. The robot starts in cell location x_0 and has observed its corresponding measurement z_{x_0}. That is, the prior data d_0 are available. Given these assumptions, we can

[2]Note that the results to follow do not rely on the grid resolution.

obtain the following result indicating that a non-extreme length-scale is a sufficient condition for adaptivity to yield positive performance advantage:

Theorem 8.2.2. If $\ell \in (0, \infty)$, $U_0^{\pi^1}(d_0) > U_0^{\pi^2}(d_0)$.

The contrapositive of Theorem 8.2.2 indicates that an extreme length-scale is a necessary condition for adaptivity to yield zero performance advantage. To prove the above result, it can be derived that the lines $\mathbb{H}[Y_{x_2^\ell} \mid d_1]$ and $\mathbb{H}[Y_{x_2^r} \mid d_1]$ have the same gradients and y-intercepts (i.e., $w_1(x_2^\ell) = w_1(x_2^r)$ and $c(d_0, x_1, x_2^\ell) = c(d_0, x_1, x_2^r)$). In contrast, the line $\mathbb{H}[Y_{x_2^f} \mid d_1]$ has a steeper gradient if the length-scale is non-extreme, which can be observed from

$$w_1(x_2^f) - w_1(x_2^r) = \frac{\sigma(1)(\sigma(\sqrt{2}) - \sigma(2))}{\sigma(0)^2 - \sigma(1)^2} = \frac{\exp\{-1.5/\ell^2\}}{1 + \frac{\rho^2 - 1}{1 - \exp\{-1/\ell^2\}}} \ . \tag{8.7}$$

that if $\ell \in (0, \infty)$, $w_1(x_2^f) > w_1(x_2^r)$. Note that $\rho = 1 + \dfrac{\sigma_n^2}{\sigma_s^2} \geq 1$. This implies the line $\mathbb{H}[Y_{x_2^f} \mid d_1]$ has to dominate the right interval of Z_{x_1} while the other line $\mathbb{H}[Y_{x_2^r} \mid d_1]$ (or $\mathbb{H}[Y_{x_2^\ell} \mid d_1]$) dominates the left. Therefore, there exists an interval, say $\mathcal{Z}_{x_1}^{[i]}$, such that $\mathbb{H}[Y_{x_2^{[i]}} \mid d_1] > \mathbb{H}[Y_{x_2^*} \mid d_1]$ for $z_{x_1} = \underline{z}_{x_1}^{[i]}$ since $\underline{z}_{x_1}^{[i]}$ cannot be a breakpoint. It follows from the last equality in (8.4) that $D(x_1, d_0) > 0$ for any x_1. Therefore, by (8.6), $U_0^{\pi^1}(d_0) > U_0^{\pi^2}(d_0)$.

The corollary below follows immediately from Theorem 8.2.1 and contrapositive of Theorem 8.2.2:

Corollary 8.2.1. $\ell = 0$ or $\ell = \infty$ iff $U_0^{\pi^1}(d_0) = U_0^{\pi^2}(d_0)$.

Fig. 8.1 shows the performance advantage of adaptivity $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0)$ with varying length-scales ℓ and signal variances σ_s^2. The noise variance σ_n^2, z_{x_0}, and μ are set to be 0.16,

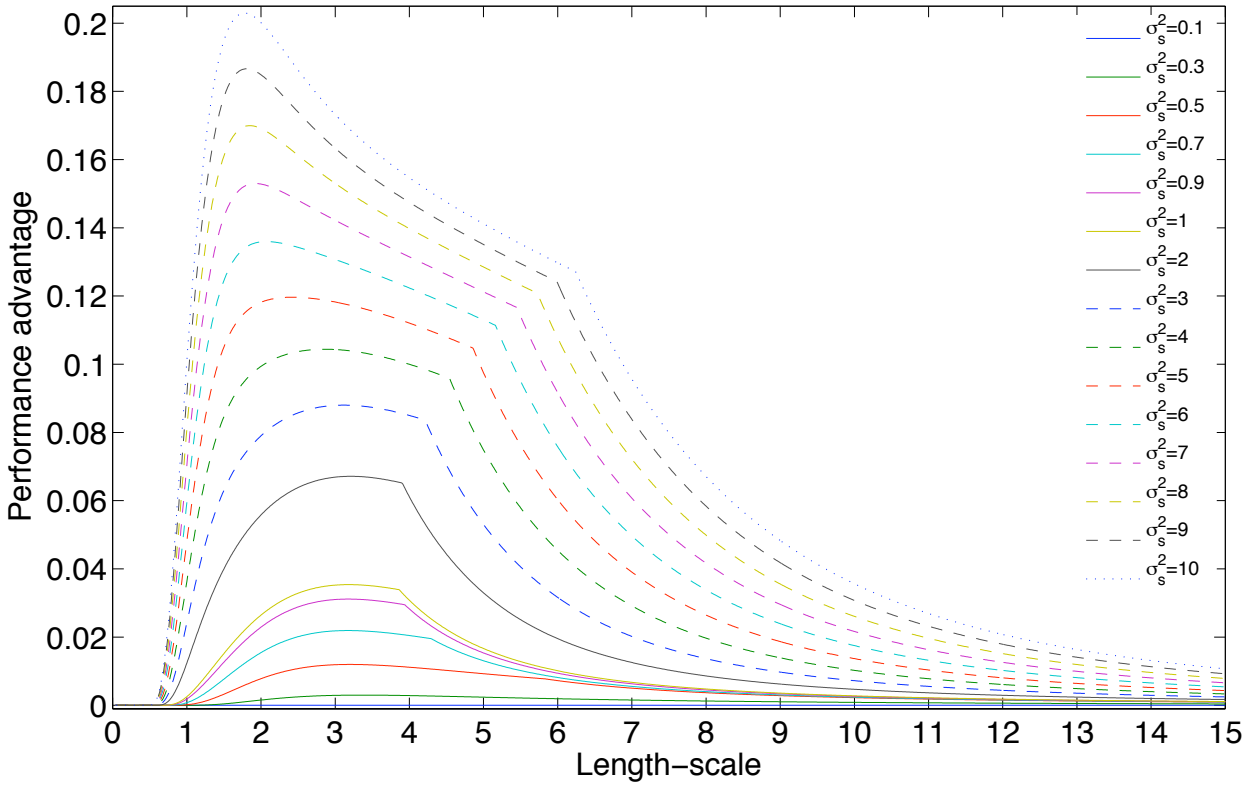


Figure 8.1: Graph of performance advantage $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0)$ vs. length-scale ℓ for varying signal variances σ_s^2.

5, and 4 respectively. It can be derived that the performance advantage of adaptivity is precisely the performance advantage that results from (and is the same for) selecting any location x_1. That is, $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0) = D(x_1^f, d_0) = D(x_1^\ell, d_0) = D(x_1^r, d_0)$. Hence, it can be evaluated in closed-form and its derivation is the same as that shown in Appendix A.15. From Fig. 8.1, we can see that the performance advantage $U_0^{\pi^1}(d_0) - U_0^{\pi^2}(d_0)$ is positive for any non-extreme length-scale, which is in agreement with Theorem 8.2.2. Increasing the signal variance σ_s^2 (i.e., decreasing noise-to-signal ratio) boosts the performance advantage at any non-extreme length-scale. For every signal variance σ_s^2, the performance advantage peaks at some length-scale. As the signal variance σ_s^2 increases, this peak shifts towards smaller length-scale ℓ and stabilizes at about $\ell = 1.8$ m. To explain this, note that the point of intersection of the lines $\mathbb{H}[Y_{x_2^f} \mid d_1]$ and $\mathbb{H}[Y_{x_2^r} \mid d_1]$ (i.e., breakpoint) has the expression

$$b \triangleq \frac{c(d_0, x_1, x_2^r) - c(d_0, x_1, x_2^f)}{w_1(x_2^f) - w_1(x_2^r)} \ .$$

We observe that, for length-scales smaller than about 2.5 m, this breakpoint is dominated by one of its terms

$$\frac{w_0(x_2^r) - w_0(x_2^f)}{w_1(x_2^f) - w_1(x_2^r)} = \rho \exp\left\{\frac{1}{2\ell^2}\right\} \ .$$

Increasing the signal variance σ_s^2 decreases ρ above, thus decreasing the breakpoint b. Consequently, this increases $\underline{p}_{x_1}^{[2]} = \frac{1}{2}\left[\operatorname{erf}\left(\frac{z_{x_1} - \mu_{Z_{x_1}|d_0}}{\sqrt{2}\sigma_{Z_{x_1}|d_0}}\right)\right]_{\mathcal{Z}_{x_1}^{[2]}=[b,\infty]} = \frac{1}{2}\left\{1 - \operatorname{erf}\left(\frac{b - \mu_{Z_{x_1}|d_0}}{\sqrt{2}\sigma_{Z_{x_1}|d_0}}\right)\right\}$, which gives a higher value of $D(x_1, d_0)$ (8.4) for small length-scales. Therefore, the peak performance advantage of adaptivity is shifted left towards smaller length-scale. This implies if the hotspots are of higher intensity, they have to be smaller in size or width for adaptivity to offer the greatest benefit.

Chapter 9

Fast Information-Theoretic Path Planning with Deterministic MDP for Active Sampling of Gaussian Process

From Section 4.3.1, we know that iMASP(1) for sampling GP (4.9) can be reduced to a non-Markovian, deterministic planning problem (Section 9.1). Due to its non-Markovian structure, the state size grows exponentially with the number of stages. Furthermore, the time complexity of evaluating each entropy-based stagewise reward in iMASP(1) depends cubically on the length of the history of observations, which limits the practical use of its approximation algorithm in *in situ* real-time, high-resolution active sampling. This latter computational difficulty also plagues the widely-used non-Markovian greedy algorithm as it is a single-staged, myopic variant of iMASP(1).

In this chapter, we will develop computationally efficient exploration strategies for sampling the GP by assuming the Markov property in iMASP(1) planning. The resulting information-theoretic path planning problem can be cast as a *deterministic Markov decision process* (DMDP) (Section 9.3). We analyze the time complexity of solving the DMDP-based path planning problem, and demonstrate analytically that it scales better than the non-

Markovian greedy algorithm with increasing number of planning stages. We also provide a theoretical guarantee on the performance of the DMDP-based policy for the case of a single robot, which, in particular, improves with decreasing spatial correlation.

Unfortunately, the performance guarantee of the DMDP-based policy cannot be generalized to the case of multiple robots unless we impose more restrictive assumptions on the GP covariance structure. However, we can obtain a similar form of performance guarantee by factoring the stagewise reward (Section 9.4), which essentially imposes a conditional independence assumption. The resulting path planning problem is therefore framed as a DMDP with factored reward (DMDP+FR). In terms of time complexity, we show analytically that it scales better than the DMDP-based algorithm with increasing number of robots.

In Section 9.5, the Markov-based algorithms are applied to the transect sampling task (Section 9.2). In particular, we investigate empirically the effects of varying spatial correlations on the mapping performance of the Markov-based policies as well as whether these Markov-based path planners are time-efficient for *in situ* real-time, high-resolution active sampling.

9.1 Deterministic Non-Markovian iMASP(1)

For sampling GP, we have learned from (4.8) that the stagewise rewards of iMASP(1) are independent of the measurements. So, each stagewise reward $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i]$ (4.8) does not have to be conditioned on the measurements $\mathbf{z}_{\mathbf{x}_{0:i}}$, and can thus be simplified to $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}]$. As a result, iMASP(1) for sampling GP (4.9) can be reduced to a non-Markovian, deterministic planning problem:

$$\begin{aligned} U_i^{\pi^1}(\mathbf{x}_{0:i}) &= \max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) \\ U_t^{\pi^1}(\mathbf{x}_{0:t}) &= \max_{\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_t,\mathbf{a}_t)} \mid \mathbf{x}_{0:t}] \end{aligned} \tag{9.1}$$

for stage $i = 0, \ldots, t-1$. Recall from Section 4.3.1 that the induced optimal policy π^1 selects observation paths with maximum entropy and is non-adaptive.

Due to the non-Markovian structure of iMASP(1) (9.1), the state size grows exponentially with the number of stages. To alleviate this computational difficulty, an anytime heuristic search algorithm called Learning Real-Time A* (Korf, 1990) is used to solve iMASP(1) approximately (Chapter 6). However, such an algorithm does not guarantee the performance of its induced policy. We have also noticed that when the action space $|\mathcal{A}(\mathbf{x}_i)|$ and the number of stages are large, it no longer produces a good policy fast enough. Even after incurring a huge amount of time and space to improve its search, its resulting policy still performs worse than the non-Markovian greedy policy (i.e., by repeatedly solving iMASP(1) with $t = 0$).

The non-Markovian structure of iMASP(1) (9.1) presents another computational problem for the anytime algorithm as well as the greedy one: the time complexity of evaluating each stagewise reward $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}]$ depends cubically on the length of the history of observations. Therefore, a limit has to be imposed on the length of the history of observations (e.g., by restricting the size of prior data or length of observation path) for higher-resolution active sampling and practical, real-time deployment of these algorithms. In Section 9.3, we do this by assuming the Markov property in iMASP(1); the resulting information-theoretic path planning problem can be cast as a *deterministic Markov decision process* (DMDP). Before discussing the DMDP-based path planning algorithms, we will first describe the exploration task that is considered in the work of this chapter.

9.2 Transect Sampling Task

Fig. 9.1 illustrates the transect sampling task that was previously introduced in (Ståhl *et al.*, 2000; Thompson and Wettergreen, 2008). The 25 m $\times$ 150 m exploration region (i.e., temperature field) is discretized into a 5×30 grid of sampling locations comprising 30 columns,

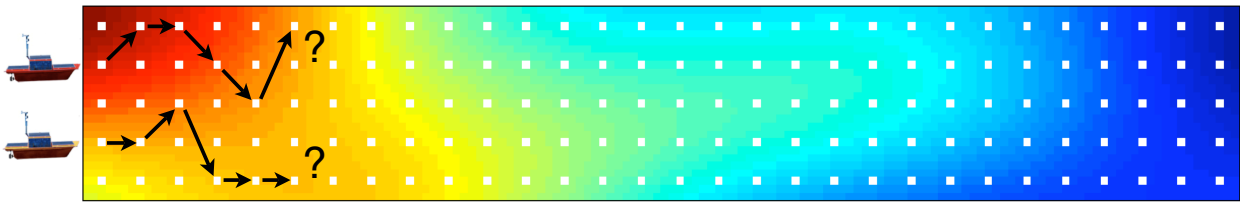


Figure 9.1: Transect sampling task over a 25 m $\times$ 150 m temperature field discretized into a 5×30 grid of sampling locations (white dots).

each of which has 5 sampling locations. It can be observed that the number of columns is much greater than the number of sampling locations in each column (i.e., number of rows). This property is assumed to be consistent with every other transect sampling region. The robots are constrained to explore forward from the leftmost to the rightmost column of the temperature field such that each robot can only sample one location per column for a total of 30 locations. So, each robot's action space given its current location consists of moving to any of the 5 locations in the adjacent column on its right. The number of robots is assumed to be smaller than or equal to the number of sampling locations per column (i.e., number of rows). We assume that an adversary chooses the starting locations of the robots in the leftmost column and the robots will only know them at the time of deployment. This assumption of unknown starting locations is realized in environments with unknown obstacles or situations when the robots do not know, prior to exploring this region, the exact locations that they will land in from exploring the previous transect sampling region or due to external forces translating them (e.g., ocean drift on the autonomous boats). The robots are allowed to end at any location in the rightmost column.

In practice, the constraint on forward exploration allows smoother motion paths and makes it easier for the robots (e.g., autonomous boats) to achieve the planned waypoints without needing complex control algorithms to perform complicated or awkward motion commands. For practical applications, during a geologic site survey (Thompson and Wettergreen, 2008), this task can be performed while the robot is en route from its current location

to a distant waypoint to collect the most "informative" data for mapping the environment between the waypoints. In monitoring of the ocean phenomena, we can think of the transect sampling region (e.g., plankton density field) drifting at a constant rate from the right to the left and the autonomous boats are tasked to explore within a line that is perpendicular to the drift.

In this transect sampling task, we assume the environmental field to vary as a realization of the Gaussian process (GP). The information-theoretic path planning algorithms considered in this chapter are therefore non-adaptive (Section 4.3.1). As a result, the observation paths can be determined prior to exploration.

9.3 Deterministic Markov Decision Process (DMDP)

By imposing the Markov assumption on iMASP(1), the resulting information-theoretic path planning problem for sampling GP can be modeled as a deterministic Markov decision process (DMDP). Specifically, the Markov property assumes each stagewise reward $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}]$ (9.1) does not have to be conditioned on the entire history of locations $\mathbf{x}_{0:i}$ but rather on the current locations $\mathbf{x}_i$ only. Hence, $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}]$ (9.1) can be approximated by $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_i]$. Imposing the Markov assumption on iMASP(1) (9.1) therefore yields the following dynamic programming equations for the DMDP-based path planning problem:

$$\begin{aligned}
\widetilde{U}_i(\mathbf{x}_i) &= \max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_i] + \widetilde{U}_{i+1}(\mathbf{x}_{i+1}) \\
&= \max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \mid \mathbf{x}_i] + \widetilde{U}_{i+1}(\tau(\mathbf{x}_i,\mathbf{a}_i)) \qquad (9.2) \\
\widetilde{U}_t(\mathbf{x}_t) &= \max_{\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_t,\mathbf{a}_t)} \mid \mathbf{x}_t]
\end{aligned}$$

for stage $i = 0, \ldots, t-1$. The optimal deterministic policy $\widetilde{\pi} = \langle \widetilde{\pi}_0(\mathbf{x}_0), \ldots, \widetilde{\pi}_t(\mathbf{x}_t) \rangle$, which is induced by solving the DMDP-based path planning problem (9.2), can be determined by

$$\begin{aligned} \widetilde{\pi}_i(\mathbf{x}_i) &= \arg\max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \mid \mathbf{x}_i] + \widetilde{U}_{i+1}(\tau(\mathbf{x}_i, \mathbf{a}_i)) \\ \widetilde{\pi}_i(\mathbf{x}_t) &= \arg\max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_t, \mathbf{a}_t)} \mid \mathbf{x}_t] \end{aligned} \tag{9.3}$$

for stage $i = 0, \ldots, t-1$. From (9.3), policy $\widetilde{\pi}$ can be used to generate optimal observation paths from all possible starting robot locations $\mathbf{x}_0$ prior to exploration because policy $\widetilde{\pi}$ is independent of the measurements $\mathbf{z}_{\mathbf{x}_{0:t}}$. As demonstrated in the experimental results (Section 9.5), policy $\widetilde{\pi}$ can be computed extremely fast especially for transect sampling tasks with unknown starting locations (Section 9.2). In contrast, the non-Markovian greedy policy incurs a considerable amount of time to be derived.

Theorem 9.3.1. Let $\mathcal{A} \triangleq \mathcal{A}(\mathbf{x}_0) = \ldots = \mathcal{A}(\mathbf{x}_t)$ and k be the number of robots. Solving the DMDP-based path planning problem (9.2) for the transect sampling task requires $\mathcal{O}(|\mathcal{A}|^2(t + k^4))$ time.

For the transect sampling task, note that the number of columns is the number of the stages plus one (i.e., $t + 2$), and the size of the joint state space within each column is equal to the size of the action space $|\mathcal{A}| = {}^r\mathrm{C}_k = \mathcal{O}(r^k)$ where r is the number of sampling locations per column (i.e., number of rows) and $k \leq r$. For each current joint state or vector of current robot locations $\mathbf{x}_i$, the time needed to evaluate the stagewise rewards $\mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \mid \mathbf{x}_i]$ (i.e., using Cholesky factorization) over all possible actions $\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)$ is $|\mathcal{A}| \times \mathcal{O}(k^4) = \mathcal{O}(|\mathcal{A}|k^4)$. Doing this over all possible current joint states within each column thus incurs $|\mathcal{A}| \times \mathcal{O}(|\mathcal{A}|k^4) = \mathcal{O}(|\mathcal{A}|^2 k^4)$ time. However, we do not have to compute these entropy-based rewards again for every column because the rewards evaluated for any one

column replicate across different columns. This computational saving is attributed primarily to the Markov assumption and, to a lesser extent, the problem structure of the transect sampling task. Propagating the optimal values from the last to the first stage takes $\mathcal{O}(|\mathcal{A}|^2 t)$ time. As a result, the time complexity of solving the DMDP-based path planning problem is $\mathcal{O}(|\mathcal{A}|^2(t+k^4))$. Though the size of the action space $|\mathcal{A}|$ is exponential in the number of robots k, the number of sampling locations per column (i.e., number of rows) r is expected to be kept small for the transect sampling task, which prevents $|\mathcal{A}|$ from growing too large.

In contrast, the time complexity of solving the non-Markovian iMASP(1) is $\mathcal{O}(|\mathcal{A}|^t t^2 k^4)$[1]. For the non-Markovian greedy algorithm, it incurs $\mathcal{O}(|\mathcal{A}|t^4k^3 + |\mathcal{A}|^2 tk^4)$ time to compute the optimal paths for all $|\mathcal{A}|$ possible choices of starting robot locations. The greedy algorithm clearly does not scale as well as the DMDP-based one with increasing number of columns (i.e., larger t), which is expected in the transect sampling task.

We will now provide a theoretical guarantee on the performance of the DMDP-based policy $\widetilde{\pi}$ vs. the iMASP(1)-based policy π^1 (9.1) for the case of 1 robot. In terms of notation, we can simplify the dynamic programming equations of the DMDP-based path planning problem (9.2) to

$$\begin{aligned} \widetilde{U}_i(x_i) &= \max_{a_i \in \mathcal{A}(x_i)} \mathbb{H}[Z_{x_{i+1}} \mid x_i] + \widetilde{U}_{i+1}(x_{i+1}) \\ \widetilde{U}_t(x_t) &= \max_{a_t \in \mathcal{A}(\mathbf{x}_t)} \mathbb{H}[Z_{x_{t+1}} \mid x_t] \ . \end{aligned} \tag{9.4}$$

In the analysis of the performance of the DMDP-based policy $\widetilde{\pi}$ below, we will assume that the GP covariance structure is defined by the squared exponential covariance function

$$\sigma_{Z_u Z_v} \stackrel{\triangle}{=} \sigma_s^2 \exp\left\{-\frac{1}{2}(u-v)^\top \mathbf{M}(u-v)\right\} + \sigma_n^2 \delta_{uv} \tag{9.5}$$

[1]Since iMASP(1) for sampling GP is the same as iMASP(n) (4.9), we can alternatively determine the time complexity of solving iMASP(n).

where σ_s^2 is the signal variance, σ_n^2 is the noise variance, $\mathbf{M} = \mathrm{diag}(\boldsymbol{\ell})^{-2}$, $\boldsymbol{\ell}$ is a vector with length-scale components ℓ_x and ℓ_y in the horizontal and vertical directions, respectively, and δ_{uv} is a Kronecker delta of value 1 if $u = v$, and 0 otherwise. For the discretization of the transect sampling region into a grid of sampling locations, let ω_x and ω_y denote the horizontal and vertical grid discretization widths (i.e., horizontal and vertical separations between adjacent sampling locations), respectively. Also, let $\ell'_x \triangleq \ell_x/\omega_x$ and $\ell'_y \triangleq \ell_y/\omega_y$ represent the normalized horizontal and vertical length-scale components, respectively.

Recall that the Markov property assumes each stagewise reward $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}]$ (9.1) can be approximated by $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_i]$ (9.4). To obtain the performance guarantee for the DMDP-based policy $\widetilde{\pi}$, we must first consider bounding the difference of these entropies ensuing from the Markov assumption:

$$\begin{aligned}\mathbb{H}[Z_{x_{i+1}} \mid x_i] - \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] &= \frac{1}{2}\log\frac{\sigma^2_{Z_{x_{i+1}}|x_i}}{\sigma^2_{Z_{x_{i+1}}|\mathbf{x}_{0:i}}} \\ &= \frac{1}{2}\log\left(1 - \frac{\sigma^2_{Z_{x_{i+1}}|x_i} - \sigma^2_{Z_{x_{i+1}}|\mathbf{x}_{0:i}}}{\sigma^2_{Z_{x_{i+1}}|x_i}}\right)^{-1} \\ &> 0\end{aligned} \tag{9.6}$$

where $\mathbf{x}_{0:i}$ is a vector concatenating $x_0, \ldots, x_i$. The following lemma bounds the variance reduction term $\sigma^2_{Z_{x_{i+1}}|x_i} - \sigma^2_{Z_{x_{i+1}}|\mathbf{x}_{0:i}}$ in (9.6):

Lemma 9.3.1. Let $\xi \triangleq \exp\left\{-\frac{1}{2\ell'^2_x}\right\}$ and $\rho \triangleq 1 + \frac{\sigma_n^2}{\sigma_s^2}$. If $\xi < \frac{\rho}{i}$,

$$0 \leq \sigma^2_{Z_{x_{i+1}}|x_i} - \sigma^2_{Z_{x_{i+1}}|\mathbf{x}_{0:i}} \leq \frac{\sigma_s^2\xi^4}{\frac{\rho}{i} - \xi}\ .$$

The proof of the above result is provided in Appendix A.16.

The next lemma is fundamental to the results on the performance of DMDP-based policy $\widetilde{\pi}$ that follow. It provides bounds on the difference of entropies $\mathbb{H}[Z_{x_{i+1}} \mid x_i] - \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}]$ arising from the Markov assumption. It follows immediately from (9.6), Lemma 9.3.1, and the following lower bound on $\sigma^2_{Z_{x_{i+1}}|x_i}$:

$$\begin{aligned} \sigma^2_{Z_{x_{i+1}}|x_i} &= \sigma^2_{Z_{x_{i+1}}} - \frac{(\sigma_{Z_{x_{i+1}} Z_{x_i}})^2}{\sigma^2_{Z_{x_i}}} \\ &\geq \sigma_s^2 + \sigma_n^2 - \sigma_s^2 \xi^2 \ . \end{aligned}$$

Lemma 9.3.2. If $\xi < \dfrac{\rho}{i}$,

$$0 \leq \mathbb{H}[Z_{x_{i+1}} \mid x_i] - \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] \leq \Delta(i)$$

where

$$\Delta(i) \triangleq \frac{1}{2} \log \left(1 - \frac{\xi^4}{(\frac{\rho}{i} - \xi)(\rho - \xi^2)} \right)^{-1} \ .$$

Remark. If $j \leq s$, $\Delta(j) \leq \Delta(s)$ for $j, s = 0, \ldots, t$.

From Lemma 9.3.2, the upper bound $\Delta(i)$ depends on the normalized length-scale ℓ'_x in the horizontal direction. As $\ell'_x \to 0^+$, $\xi \to 0^+$ and consequently, $\Delta(i) \to 0^+$. This means when the horizontal correlation tends to zero, the difference of the entropies $\mathbb{H}[Z_{x_{i+1}} \mid x_i]$ and $\mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}]$ ensuing from the Markov assumption disappears. The upper bound $\Delta(i)$ also depends on the noise-to-signal ratio σ_n^2/σ_s^2 through ρ. Raising the noise-to-signal ratio decreases the difference of the entropies $\mathbb{H}[Z_{x_{i+1}} \mid x_i]$ and $\mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}]$. Lastly, the value of i indicates the length of history of observations. The remark in Lemma 9.3.2 tells us that a shorter length produces a smaller upper bound $\Delta(i)$, and hence a smaller difference of the

entropies $\mathbb{H}[Z_{x_{i+1}} \mid x_i]$ and $\mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}]$. One limitation with using this upper bound $\Delta(i)$ is that the condition $\xi < \rho/i$ has to be satisfied. Hence, it cannot be used to analyze the case of extremely large horizontal correlation.

The following theorem uses the induced optimal value $\widetilde{U}_0(x_0)$ from solving the DMDP-based path planning problem (9.4) to bound the largest entropy of observation paths $U_0^{\pi^1}(x_0)$ achieved by policy π^1 from solving iMASP(1) (9.1):

Theorem 9.3.2. Let $\epsilon(i) \triangleq \sum_{s=i}^{t} \Delta(s) \leq (t-i+1)\Delta(t)$. Then, $\widetilde{U}_i(x_i) - \epsilon(i) \leq U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq \widetilde{U}_i(x_i)$ for $i = 0, \ldots, t$.

The proof of the above result uses Lemma 9.3.2 and is provided in Appendix A.17. Since the error bound $\epsilon(i)$ depends on the sum of $\Delta(s)$'s, we can rely upon the observations on $\Delta(s)$ (see paragraph just after Lemma 9.3.2) to know how the error bound $\epsilon(i)$ can be improved. In essence, smaller horizontal correlation, larger noise-to-signal ratio, and shorter length of history of observations improve the error bound $\epsilon(i)$.

In the result below, the DMDP-based policy $\widetilde{\pi}$ is guaranteed to achieve an entropy of observation paths $U_0^{\widetilde{\pi}}(x_0)$ that is not more than $\sum_{s=0}^{t} \Delta(s)$ from the largest entropy of observation paths $U_0^{\pi^1}(x_0)$ achieved by policy π^1.

Theorem 9.3.3. Define the entropy of observation paths achieved by policy π with the following value functions

$$\begin{aligned} U_i^{\pi}(\mathbf{x}_{0:i}) &= \mathbb{H}[Z_{\tau(x_i, \pi_i(\mathbf{x}_{0:i}))} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\pi}(\mathbf{x}_{0:i+1}) \\ U_t^{\pi}(\mathbf{x}_{0:t}) &= \mathbb{H}[Z_{\tau(x_t, \pi_i(\mathbf{x}_{0:t}))} \mid \mathbf{x}_{0:t}] \end{aligned} \tag{9.7}$$

for stage $i = 0, \ldots, t-1$. Let $\epsilon \triangleq \sum_{s=0}^{t} \Delta(s) \leq (t+1)\Delta(t)$. Then, policy $\widetilde{\pi}$ is ϵ-optimal for achieving the entropy criterion. That is, $U_0^{\pi^1}(x_0) - U_0^{\widetilde{\pi}}(x_0) \leq \epsilon$.

The proof of the above result uses Lemma 9.3.2 and is provided in Appendix A.18. Again, since the error bound ϵ depends on the sum of $\Delta(s)$'s, we can rely upon the observations on $\Delta(s)$ (see paragraph just after Lemma 9.3.2) to understand how the error bound ϵ can be improved. In essence, smaller horizontal correlation, larger noise-to-signal ratio, and shorter length of history of observations improve the error bound ϵ.

9.4 Deterministic Markov Decision Process with Factored Reward (DMDP+FR)

Unfortunately, the performance guarantee of the DMDP-based policy $\widetilde{\pi}$ cannot be generalized to the case of multiple robots unless we impose more restrictive assumptions on the GP covariance structure such as the absence of suppressor variables (i.e., $|\sigma_{Z_j Z_k|\mathbf{x}_i}| \leq \sigma_{Z_j Z_k}$). However, we can provide a similar form of performance guarantee by factoring the stagewise reward. Specifically, the stagewise reward $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_i]$ (9.2) is factored into a sum of k additive rewards $\sum_{m=1}^{k} \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}]$ where we define $\mathbf{x}_i^{[m]}$ and $\mathbf{x}_{i+1}^{[m]}$ to be the m-th components of the vectors of current robot locations $\mathbf{x}_i$ and next robot locations $\mathbf{x}_{i+1}$, respectively. So, each random measurement $Z_{\mathbf{x}_{i+1}^{[m]}}$ is assumed to be conditionally independent of every other random measurement $Z_{\mathbf{x}_{i+1}^{[j]}}$ for $j \neq m$. Factoring the stagewise reward $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_i]$ in the DMDP-based path planning problem (9.2) therefore yields the following dynamic programming equations for the DMDP+FR-based problem:

$$\begin{aligned} \widehat{U}_i(\mathbf{x}_i) &= \max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \sum_{m=1}^{k} \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}] + \widehat{U}_{i+1}(\mathbf{x}_{i+1}) \\ \widehat{U}_t(\mathbf{x}_t) &= \max_{\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)} \sum_{m=1}^{k} \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_t^{[m]}] \end{aligned} \tag{9.8}$$

for stage $i = 0, \ldots, t-1$. The optimal deterministic policy $\widehat{\pi} = \langle \widehat{\pi}_0(\mathbf{x}_0), \ldots, \widehat{\pi}_t(\mathbf{x}_t) \rangle$ is induced by solving the DMDP+FR-based path planning problem (9.8).

Theorem 9.4.1. Solving the DMDP+FR-based path planning problem (9.8) for the transect sampling task requires $\mathcal{O}(|\mathcal{A}|^2(t+k))$ time.

The time needed to evaluate the entropy-based factored reward $\mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}]$ for each pair of current and next location components (i.e., respectively, $\mathbf{x}_i^{[m]}$ and $\mathbf{x}_{i+1}^{[m]}$) is $\mathcal{O}(1)$. Summing them over k pairs of location components thus incurs $\mathcal{O}(k)$ time. Doing this over all pairs of current and next joint states (i.e., respectively, $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$) within each column therefore takes $\mathcal{O}(|\mathcal{A}|^2 k)$ time[2]. Similar to that of DMDP, we do not have to compute these entropy-based rewards again for every column because the rewards evaluated for any one column are the same across different columns. Propagating the optimal values from the last to the first stage requires $\mathcal{O}(|\mathcal{A}|^2 t)$ time. As a result, the time complexity of solving the DMDP+FR-based path planning problem is $\mathcal{O}(|\mathcal{A}|^2(t+k))$. It can be observed that the DMDP+FR-based dynamic programming algorithm scales better than the DMDP-based one with increasing number of robots k.

We will now provide a theoretical guarantee on the performance of the DMDP+FR-based policy $\widehat{\pi}$ vs. the iMASP(1)-based policy π^1 (9.1) for the case of multiple robots (i.e., $k \geq 1$). We will assume here that the normalized horizontal and vertical length-scale components are equal (i.e., $\ell' \triangleq \ell'_x = \ell'_y$).

To obtain the performance guarantee for the DMDP+FR-based policy $\widetilde{\pi}$, we must first consider bounding the difference between $\sum_{m=1}^{k} \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}]$ and $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}]$ ensuing

[2]This is possible if we order the location components in every joint state $\mathbf{x}_i$ based on their corresponding row numbers, rather than through the robot id's. While this reduces the size of the action space, it does not decrease the induced optimal value from solving iMASP(1) or change its optimal policy π^1 since the order of the location components in a joint state does not affect the corresponding evaluated posterior joint entropy. It may, however, decrease the achievable optimal value from solving DMDP+FR.

from the Markov assumption and factored reward:

$$\begin{aligned}
&\sum_{m=1}^{k} \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}] - \mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}] \\
&= \sum_{m=1}^{k} \left(\mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}] - \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_{0:i}, \mathbf{x}_{i+1}^{[1:m-1]}] \right) \\
&= \frac{1}{2} \sum_{m=1}^{k} \left(\log \frac{\sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_i^{[m]}}}{\sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_{0:i}, \mathbf{x}_{i+1}^{[1:m-1]}}} \right) \\
&= \frac{1}{2} \sum_{m=1}^{k} \log \left(1 - \frac{\sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_i^{[m]}} - \sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_{0:i}, \mathbf{x}_{i+1}^{[1:m-1]}}}{\sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_i^{[m]}}} \right)^{-1} \\
&> 0
\end{aligned} \tag{9.9}$$

where $\mathbf{x}_{i+1}^{[1:m-1]}$ denotes a vector concatenating $\mathbf{x}_{i+1}^{[1]}, \ldots, \mathbf{x}_{i+1}^{[m-1]}$. Similar to Lemma 9.3.1, the following result bounds the variance reduction term $\sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_i^{[m]}} - \sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_{0:i}, \mathbf{x}_{i+1}^{[1:m-1]}}$ in (9.9):

Lemma 9.4.1. Let $\xi \triangleq \exp\left\{-\frac{1}{2\ell'^2}\right\}$ and $\rho \triangleq 1 + \frac{\sigma_n^2}{\sigma_s^2}$. If $\xi < \frac{\rho}{ki + 2(k-1)}$,

$$0 \leq \sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_i^{[m]}} - \sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_{0:i}, \mathbf{x}_{i+1}^{[1:m-1]}} \leq \frac{\sigma_s^2 \left(\frac{ki}{ki+2(k-1)} \xi^4 + \frac{2(k-1)}{ki+2(k-1)} \xi^2 \right)}{\frac{\rho}{ki+2(k-1)} - \xi} \leq \frac{\sigma_s^2 \xi^2}{\frac{\rho}{k(i+2)-2} - \xi}$$

for $m = 1, \ldots, k$.

Remarks.

1. When $k = 1$, the tighter upper bound reduces to that of Lemma 9.3.1.

2. For the multi-robot case here, ξ is defined in terms of the same normalized length-scale component in both horizontal and vertical directions (i.e., $\ell' \triangleq \ell'_x = \ell'_y$) instead

of just the normalized horizontal length-scale component ℓ'_x for the single-robot case in Lemma 9.3.1.

The proof of the above result is similar to that of Lemma 9.3.1. Let $\mathbf{x}_i^{[-m]}$ denote the vector of current robot locations $\mathbf{x}_i$ without the component $\mathbf{x}_i^{[m]}$. In comparison to the upper bound of Lemma 9.3.1, the numerator of the tighter upper bound comprises a convex combination of ξ^4 and ξ^2 terms; the former ξ^4 term is due to the Markov assumption while the latter ξ^2 term arises from the conditional independence assumption. The larger ξ^2 term is due to the presence of multiple robots residing at location components of $\mathbf{x}_i^{[-m]}$ and $\mathbf{x}_{i+1}^{[1:m-1]}$, which are close to the locations $\mathbf{x}_i^{[m]}$ and $\mathbf{x}_{i+1}^{[m]}$.

Similar to Lemma 9.3.2, the next lemma provides bounds on the difference between $\sum_{m=1}^{k} \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}]$ and $\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}]$ arising from the Markov assumption and factored reward, which is crucial to proving the results on the performance of DMDP+FR-based policy $\widehat{\pi}$ that follow. It follows immediately from (9.9), Lemma 9.4.1, and a lower bound on $\sigma^2_{Z_{\mathbf{x}_{i+1}^{[m]}} | \mathbf{x}_i^{[m]}}$:

Lemma 9.4.2. If $\xi < \dfrac{\rho}{k(i+2)-2}$,

$$0 \leq \sum_{m=1}^{k} \mathbb{H}[Z_{\mathbf{x}_{i+1}^{[m]}} \mid \mathbf{x}_i^{[m]}] - \mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid \mathbf{x}_{0:i}] \leq \Delta_1(i) \leq \Delta_2(i)$$

where

$$\Delta_1(i) \triangleq \frac{k}{2} \log \left(1 - \frac{\left(\frac{ki}{ki+2(k-1)} \xi^4 + \frac{2(k-1)}{ki+2(k-1)} \xi^2 \right)}{(\frac{\rho}{ki+2(k-1)} - \xi)(\rho - \xi^2)} \right)^{-1}$$

and

$$\Delta_2(i) \triangleq \frac{k}{2} \log \left(1 - \frac{\xi^2}{(\frac{\rho}{k(i+2)-2} - \xi)(\rho - \xi^2)} \right)^{-1} .$$

Remark. If $j \leq s$, $\Delta_2(j) \leq \Delta_2(s)$ for $j, s = 0, \ldots, t$.

To interpret the upper bounds $\Delta_1(i)$ and $\Delta_2(i)$, we can rely upon the observations on $\Delta(s)$ (see paragraph just after Lemma 9.3.2) to know how the upper bounds can be improved. In particular, smaller length-scale, larger noise-to-signal ratio, and shorter length of history of observations improve the upper bounds. Additionally, it can be observed that decreasing the number of robots k reduces $\Delta_2(i)$ and allows the sufficient condition $\xi < \rho/(k(i+2)-2)$ to be more easily satisfied.

Similar to Theorem 9.3.2, the following theorem uses the induced optimal value $\widehat{U}_0(\mathbf{x}_0)$ from solving the DMDP+FP-based path planning problem (9.8) to bound the largest entropy of observation paths $U_0^{\pi^1}(\mathbf{x}_0)$ achieved by policy π^1 from solving iMASP(1) (9.1):

Theorem 9.4.2. Let $\epsilon_1(i) \triangleq \sum_{s=i}^{t} \Delta_1(s)$ and $\epsilon_2(i) \triangleq \sum_{s=i}^{t} \Delta_2(s)$. Then, $\widehat{U}_i(\mathbf{x}_i) - \epsilon_2(i) \leq \widehat{U}_i(\mathbf{x}_i) - \epsilon_1(i) \leq U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq \widehat{U}_i(\mathbf{x}_i)$ for $i = 0, \ldots, t$.

The proof of the above result uses Lemma 9.4.2 and is similar to that of Theorem 9.3.2. Since the error bounds $\epsilon_1(i)$ and $\epsilon_2(i)$ depend, respectively, on the sum of $\Delta_1(s)$'s and $\Delta_2(s)$'s, we can rely upon the observations on $\Delta_1(s)$ and $\Delta_2(s)$ (see paragraph just after Lemma 9.4.2) to understand how the error bounds $\epsilon_1(i)$ and $\epsilon_2(i)$ can be improved. In essence, smaller length-scale, larger noise-to-signal ratio, shorter length of history of observations, and smaller number of robots improve the error bounds $\epsilon_1(i)$ and $\epsilon_2(i)$.

Similar to Theorem 9.3.3, the result below guarantees the DMDP+FP-based policy $\widehat{\pi}$ to achieve an entropy of observation paths $U_0^{\widehat{\pi}}(\mathbf{x}_0)$ that is not more than $\sum_{s=0}^{t} \Delta_1(s)$ ($\sum_{s=0}^{t} \Delta_2(s)$) from the largest entropy of observation paths $U_0^{\pi^1}(\mathbf{x}_0)$ achieved by policy π^1.

Theorem 9.4.3. Let $\epsilon_1 \triangleq \sum_{s=0}^{t} \Delta_1(s)$ and $\epsilon_2 \triangleq \sum_{s=0}^{t} \Delta_2(s)$. Then, policy $\widehat{\pi}$ is ϵ_1-optimal (ϵ_2-optimal) for achieving the entropy criterion. That is, $U_0^{\pi^1}(\mathbf{x}_0) - U_0^{\widehat{\pi}}(\mathbf{x}_0) \leq \epsilon_1 \leq \epsilon_2$.

The proof of the above result uses Lemma 9.4.2 and is similar to that of Theorem 9.3.3. Again, since the error bounds ϵ_1 and ϵ_2 depend, respectively, on the sum of $\Delta_1(s)$'s and $\Delta_2(s)$'s, we can rely upon the observations on $\Delta_1(s)$ and $\Delta_2(s)$ (see paragraph just after Lemma 9.4.2) to understand how the error bounds ϵ_1 and ϵ_2 can be improved. In essence, smaller length-scale, larger noise-to-signal ratio, shorter length of history of observations, and smaller number of robots improve the error bounds ϵ_1 and ϵ_2.

9.5 Experiments and Discussion

This section presents empirical evaluations of the induced Markov-based optimal policies $\widetilde{\pi}$ and $\widehat{\pi}$ from, respectively, solving DMDP and DMDP+FR on the temperature field data of Panther Hollow Lake in Pittsburgh, PA spanning 25 m $\times$150 m. The exploration region is discretized into a 5×30 grid of sampling locations as shown in Fig. 9.2. The setup of the transect sampling task has been described in Section 9.2. Using maximum likelihood estimation, the learned horizontal and vertical length-scale hyperparameters (i.e., respectively, ℓ_x and ℓ_y) are 40.45 m and 16.00 m, respectively.

The performances of the DMDP-based policy $\widetilde{\pi}$ and the DMDP+FR-based policy $\widehat{\pi}$ are compared with that of the non-Markovian greedy policy (i.e., by repeatedly solving iMASP(1) with $t = 0$) denoted by π_G. The non-Markovian policy π^1 is initially included for comparison and has to be derived approximately using an anytime heuristic search algorithm called Learning Real-Time A*. However, we have noticed through experiments that it no longer produces a good policy fast enough due to the large action space $|\mathcal{A}(\mathbf{x}_i)|$ and number

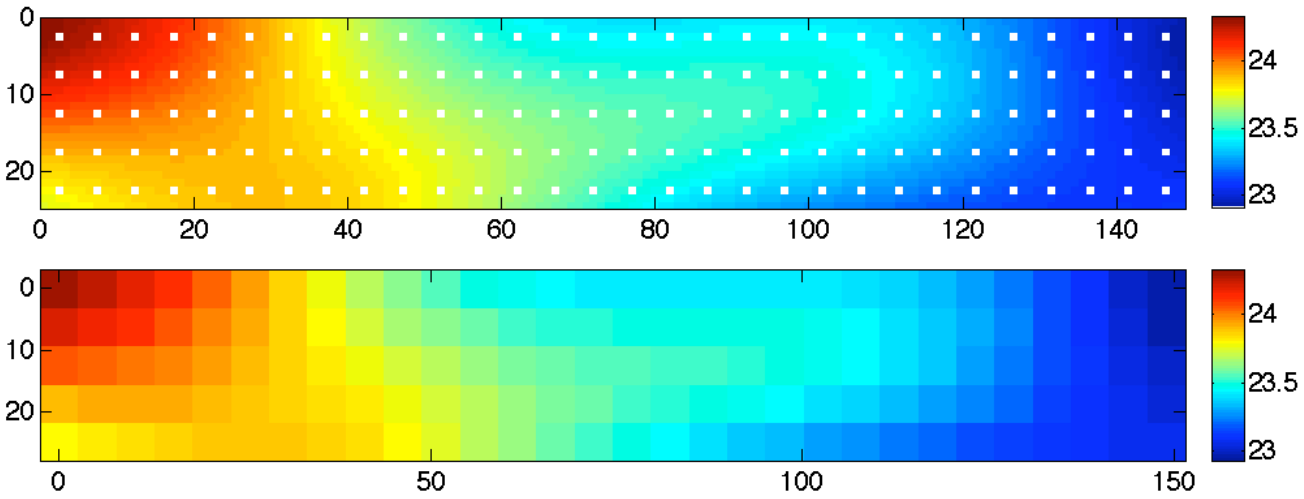


Figure 9.2: Temperature field discretized into a 5×30 grid of sampling locations (white dots).

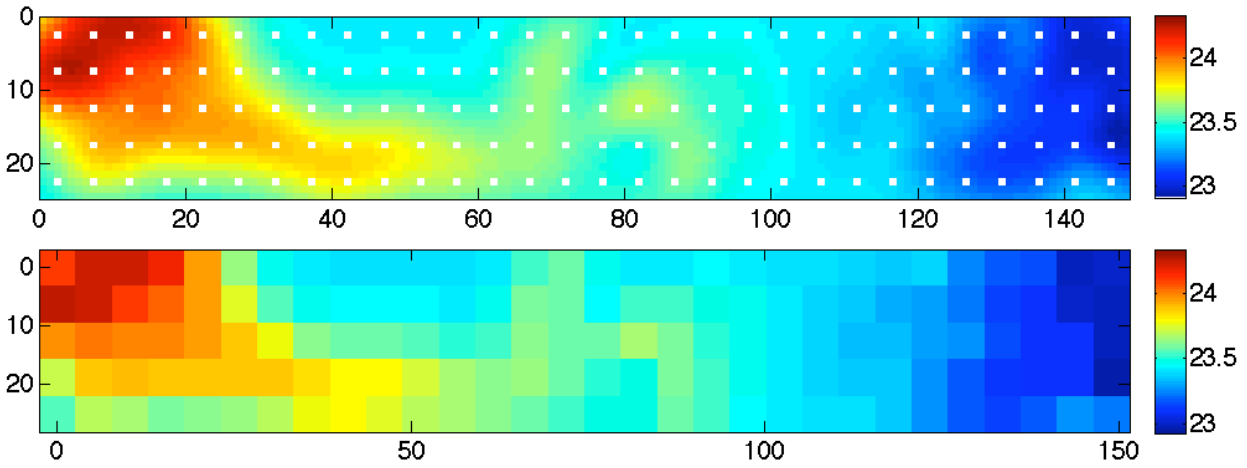
of stages involved in the transect sampling task. Even after incurring a huge amount of time and space to improve its search, its resulting policy still performs worse than the non-Markovian greedy policy π_G. Hence, it is excluded from comparison.

9.5.1 Performance metrics

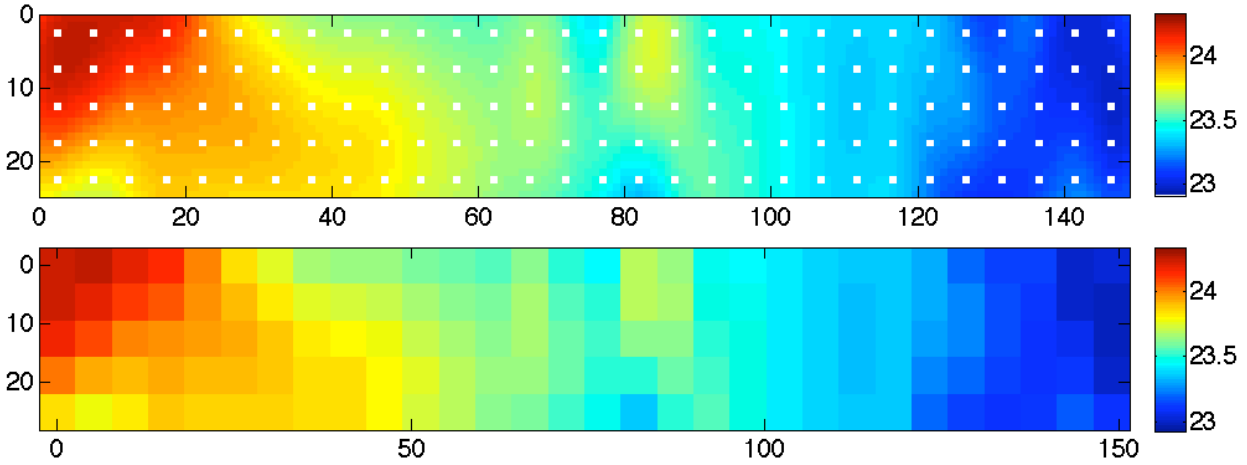
The ENT and ERR metrics described in Section 6.1 are also used here to evaluate the performance of the policies. We will sometimes use ENT(π) and ERR(π) to, respectively, denote the posterior map entropy and the mean-squared relative error achieved by policy π. Additionally, we will consider the time taken to derive each policy as a performance metric.

9.5.2 Test results

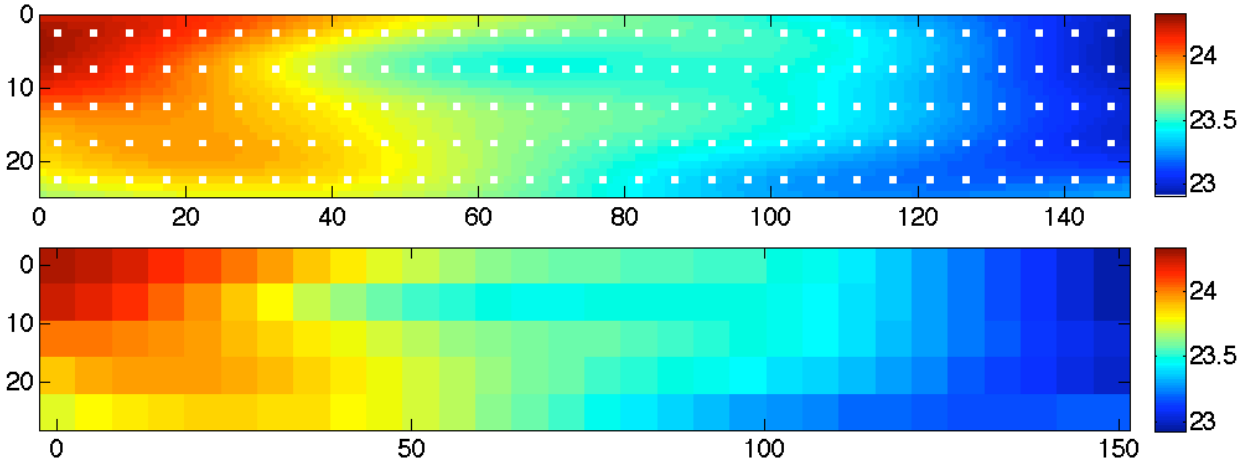
We will first investigate the effects of varying spatial correlations (in particular, varying length-scales) on the ENT and ERR performance of the evaluated policies. By reducing the horizontal and/or vertical length-scales of the original temperature field, the following



(a) Temperature field 1: $\ell_x = 5.00$ m, $\ell_y = 5.00$ m.



(b) Temperature field 2: $\ell_x = 5.00$ m, $\ell_y = 16.00$ m..



(c) Temperature field 3: $\ell_x = 40.45$ m, $\ell_y = 5.00$ m.

Figure 9.3: Temperature fields with varying horizontal and vertical length-scales.

modified fields are obtained:

Temperature field	ℓ_x (m)	ℓ_y (m)
1 (Fig. 9.3a)	5.00	5.00
2 (Fig. 9.3b)	5.00	16.00
3 (Fig. 9.3c)	40.45	5.00
4 (Fig. 9.2a)	40.45	16.00

Note that the original temperature field is field 4. To produce the modified fields 1, 2, and 3, we fix the horizontal and vertical length-scales, and use the temperature data to learn the remaining hyperparameters (i.e., signal and noise variances) through maximum likelihood estimation.

Figs. 9.4, 9.6, and 9.8 show the results of the mean ENT and ERR performance of the tested policies (i.e., averaged over all possible starting robot locations) with varying length-scales and robot team sizes. Other than the ERR performance of the policies $\widetilde{\pi}$ and $\widehat{\pi}$ for the 1-robot case, we can observe decreasing ENT and ERR for all policies with increasing length-scales due to increasing spatial correlation between measurements, thus resulting in decreasing map uncertainty.

For the case of 1 robot (Fig. 9.4), note that the DMDP-based policy $\widetilde{\pi}$ is equivalent to the DMDP+FR-based policy $\widehat{\pi}$. The observations for the 1-robot case are as follows:

1. When the vertical length-scale is kept constant (i.e., either at $\ell_y = 5$ m or at $\ell_y = 16$ m), decreasing the horizontal length-scale from $\ell_x = 40.45$ m to $\ell_x = 5$ m (i.e., either from field 3 to field 1 or from field 4 to field 2) reduces the difference in ENT and ERR performance between a Markov-based policy and the non-Markovian greedy policy π_G. This agrees with the decreasing error bound ϵ of the ϵ-optimal policy $\widetilde{\pi}$ for achieving the entropy criterion due to smaller horizontal length-scale ℓ_x (Theorem 9.3.3) even though this theoretical result holds with respect to the non-Markovian policy π_1

rather than the policy π_G. Intuitively, since the horizontal correlation is small, the non-Markovian greedy policy π_G loses its advantage of being able to exploit a large horizontal correlation.

2. The differences in ENT performance between a Markov-based policy and the non-Markovian greedy policy π_G are small for all temperature fields except for field 3 with large horizontal length-scale ℓ_x and small vertical length-scale ℓ_y: due to the Markov assumption, a Markov-based policy is not capable of exploiting the large horizontal correlation to spread out the sampled locations in each row and distribute them evenly across rows. Beyond the starting column, locations in rows 2 to 4 are therefore not sampled as shown in Fig. 9.5a, thus resulting in a higher ENT. In contrast, the non-Markovian greedy policy π_G is capable of doing so as shown in Fig. 9.5b.

3. For field 1 with small horizontal and vertical length-scales, a Markov-based policy achieves slightly lower ENT than the non-Markovian greedy policy π_G: the locations sampled by these policies are similar to those in Fig. 9.5. It can be observed that the total area of unsampled grid cells is the same for both policies, but the area corresponding to the Markov-based policy is not riddled by sampled grid cells. Consequently, the prior joint entropy of the measurements at unobserved locations is smaller for the Markov-based policy than for the non-Markovian greedy policy π_G. Since the horizontal and vertical correlations are both small, the entropy reduction due to the sampled locations is small for both policies. The resulting posterior map entropy is therefore lower for the Markov-based policy.

4. When the horizontal length-scale is kept constant (i.e., either at $\ell_x = 5$ m or at $\ell_x = 40.45$ m), decreasing the vertical length-scale from $\ell_y = 16$ m to $\ell_y = 5$ m (i.e., either from field 2 to field 1 or from field 4 to field 3) raises the difference in ERR performance between a Markov-based policy and the non-Markovian greedy policy π_G:

with small vertical correlation, the measurements at unobserved locations in rows 2 to 4 for the Markov-based policy (Fig. 9.5a) can no longer be predicted well, thus incurring larger ERRs.

For the case of more than 1 robot (Figs. 9.6 and 9.8), the DMDP-based policy $\widetilde{\pi}$ and the DMDP+FR-based policy $\widehat{\pi}$ are no longer equivalent. It can be observed from Figs. 9.6 and 9.8 that when the large horizontal and vertical length-scales of field 4 reduce to the small length-scales of field 1 (i.e., $\ell \triangleq \ell_x = \ell_y = 5$ m), the difference in ENT performance between the DMDP+FR-based policy $\widehat{\pi}$ and the non-Markovian greedy policy π_G decreases. This decrease is also observed for field 1 with large horizontal and vertical length-scales and field 4 with small horizontal and vertical length-scales when the number of robots decreases from 3 to 1. These observations agree with the decreasing error bound ϵ_1 (ϵ_2) of the ϵ_1-optimal (ϵ_2-optimal) policy $\widehat{\pi}$ for achieving the entropy criterion due to smaller length-scale ℓ and smaller number of robots k (Theorem 9.4.3) even though this theoretical result holds with respect to the non-Markovian policy π_1 rather than the policy π_G.

For the case of 2 robots (Fig. 9.6), the observations are as follows:

1. The differences in ENT performance between the DMDP+FR-based policy $\widehat{\pi}$ and the non-Markovian greedy policy π_G are small for all temperature fields except for field 2 with small horizontal length-scale ℓ_x and large vertical length-scale ℓ_y: due to the conditional independence assumption (Section 9.4), the DMDP+FR-based policy $\widehat{\pi}$ is not capable of exploiting the large vertical correlation to spread out the sampled locations in each column (Fig. 9.7a), thus incurring a higher ENT. In contrast, the DMDP-based policy $\widetilde{\pi}$ and the non-Markovian greedy policy π_G are capable of doing so (Fig. 9.7b). However, the DMDP+FR-based policy $\widehat{\pi}$ achieves slightly lower ERR than the DMDP-based policy $\widetilde{\pi}$ and the non-Markovian greedy policy π_G: it can be observed from Fig. 9.7b that beyond the starting column, the policies $\widetilde{\pi}$ and π_G do not sample locations in rows 2 to 4, thus resulting in a higher ERR.

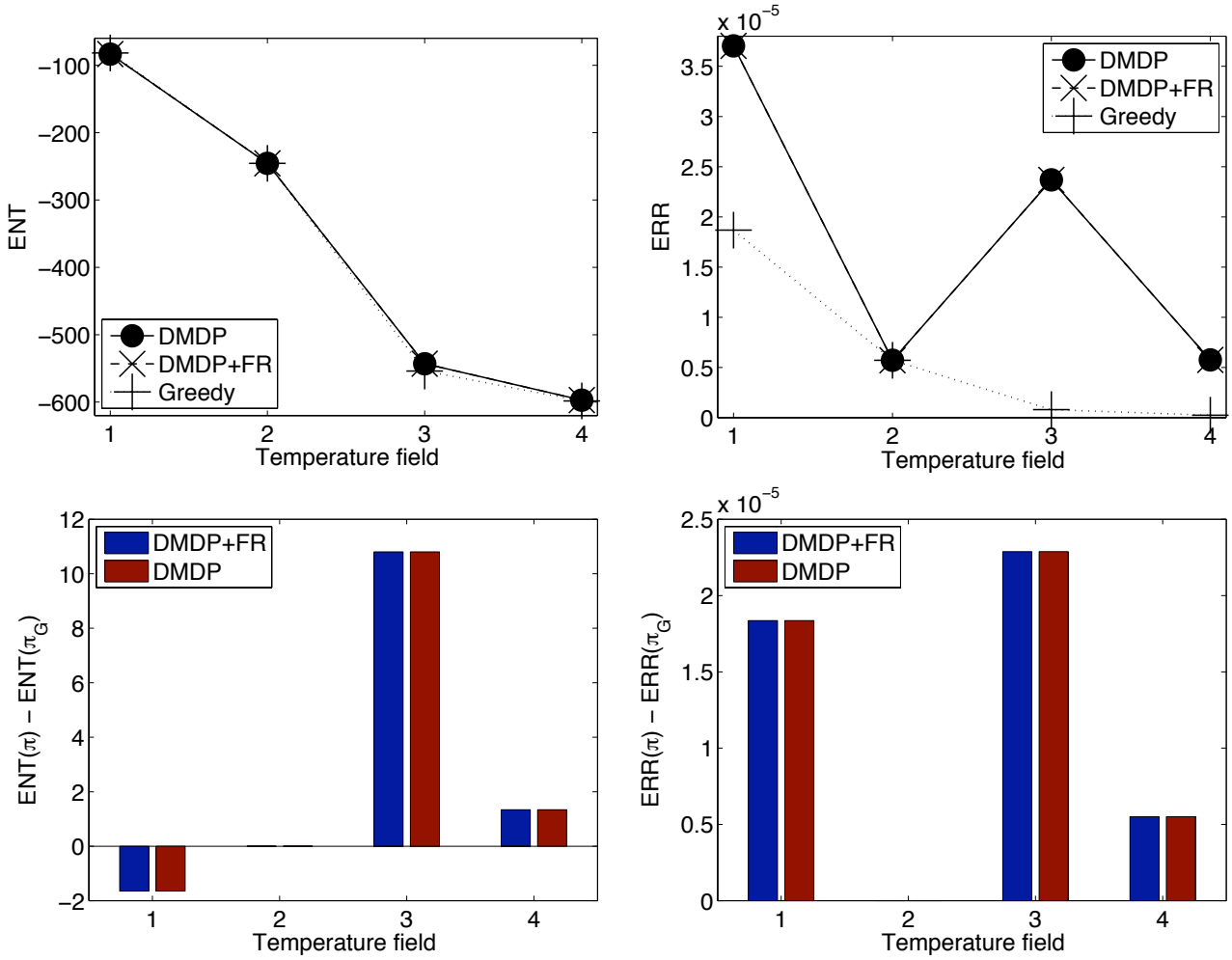
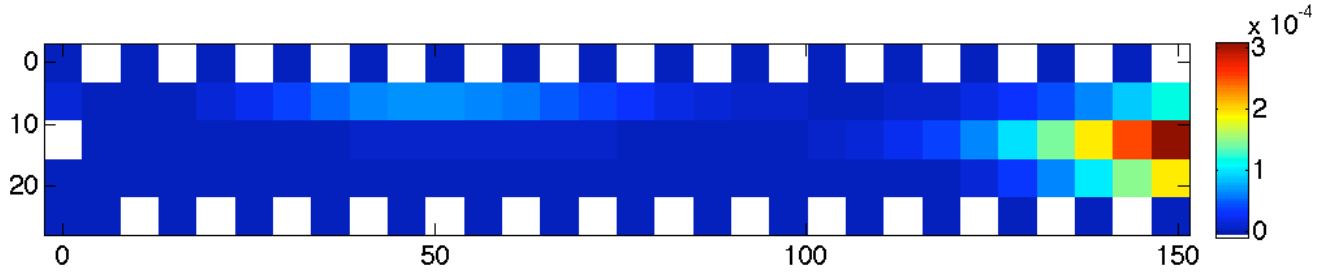
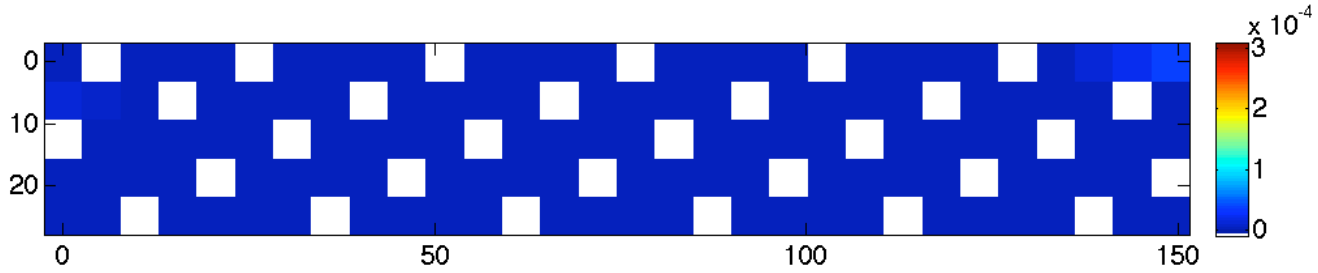


Figure 9.4: Performance comparison of DMDP-based, DMDP+FR-based, and non-Markovian greedy policies for the case of 1 robot.



(a)



(b)

Figure 9.5: Squared relative error maps for field 3 with 1 robot showing white grid cells sampled by the (a) DMDP-based policy $\widetilde{\pi}$ and (b) non-Markovian greedy policy π_G.

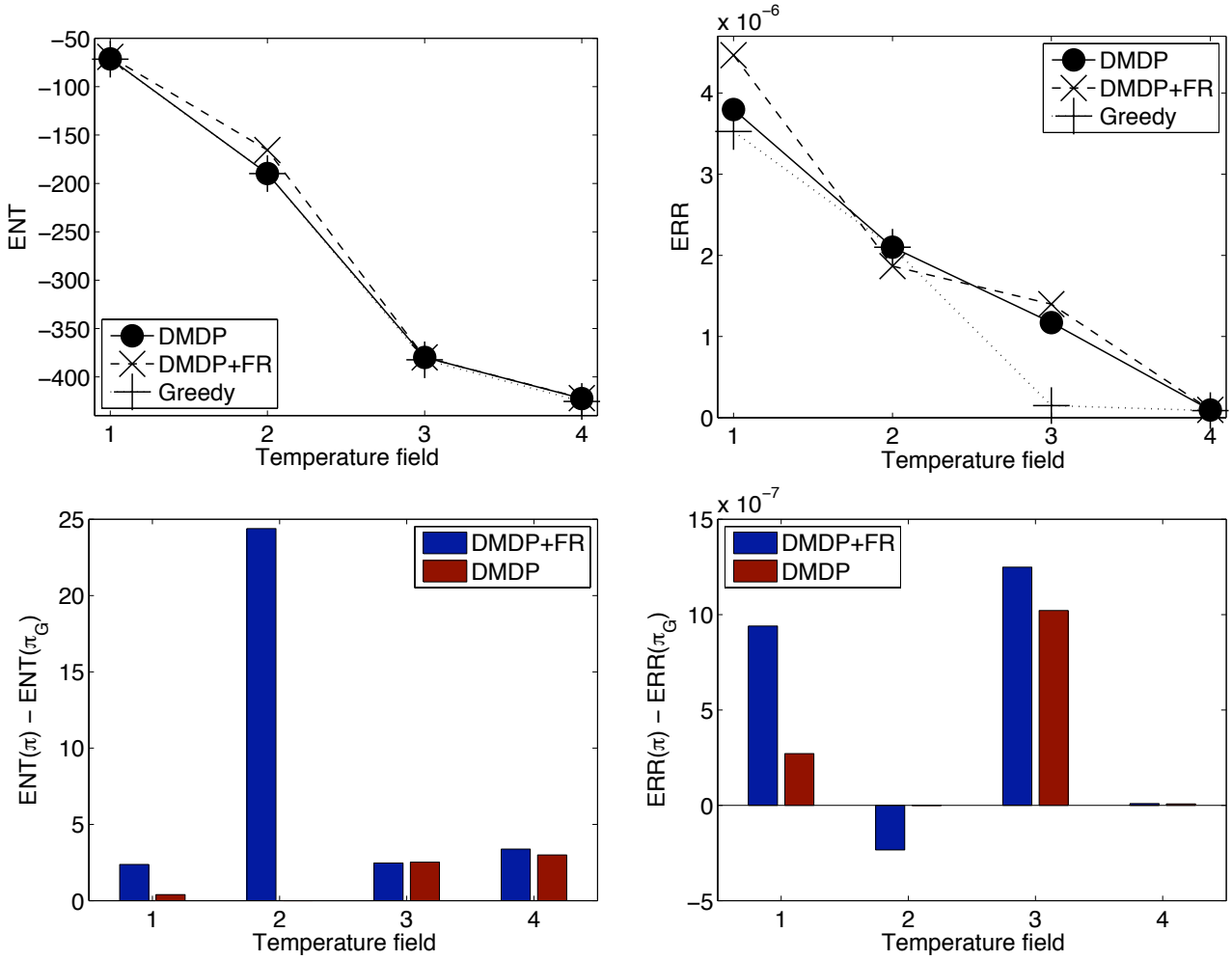
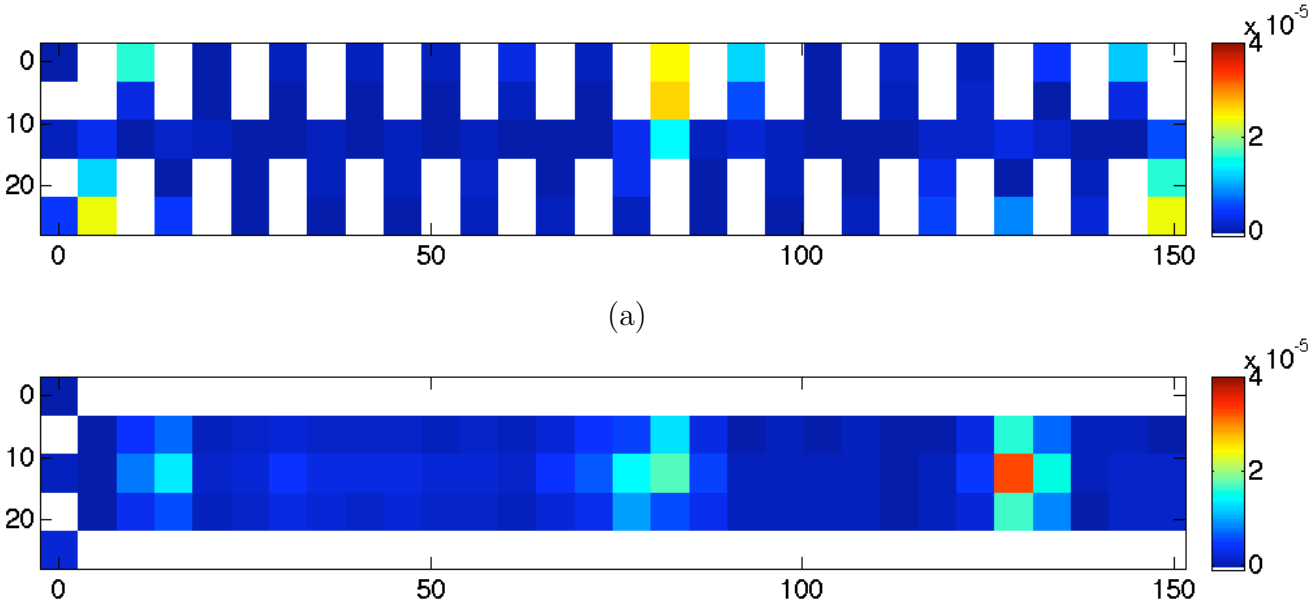


Figure 9.6: Performance comparison of DMDP-based, DMDP+FR-based, and non-Markovian greedy policies for the case of 2 robots.



(a)

(b)

Figure 9.7: Squared relative error maps for field 2 with 2 robots showing white grid cells sampled by the (a) DMDP+FR-based policy $\widehat{\pi}$, and (b) DMDP-based policy $\widetilde{\pi}$ and non-Markovian greedy policy π_G.

2. The differences in ENT performance between the DMDP-based policy $\widetilde{\pi}$ and the non-Markovian greedy policy π_G are small for all temperature fields. In particular, when the vertical length-scale is kept constant (i.e., either at $\ell_y = 5$ m or at $\ell_y = 16$ m), decreasing the horizontal length-scale from $\ell_x = 40.45$ m to $\ell_x = 5$ m (i.e., either from field 3 to field 1 or from field 4 to field 2) reduces the difference in ENT performance between policies $\widetilde{\pi}$ and π_G: this is explained in the first observation for the 1-robot case.

3. When the vertical length-scale is kept constant (i.e., either at $\ell_y = 5$ m or at $\ell_y = 16$ m), decreasing the horizontal length-scale from $\ell_x = 40.45$ m to $\ell_x = 5$ m (i.e., either from field 3 to field 1 or from field 4 to field 2) reduces the difference in ERR performance between a Markov-based policy and the non-Markovian greedy policy π_G: this is explained in the first observation for the 1-robot case.

4. When the horizontal length-scale is kept constant (i.e., either at $\ell_x = 5$ m or at $\ell_x = 40.45$ m), decreasing the vertical length-scale from $\ell_y = 16$ m to $\ell_y = 5$ m (i.e., either from field 2 to field 1 or from field 4 to field 3) raises the difference in ERR performance between a Markov-based policy and the non-Markovian greedy policy π_G: with small vertical correlation, the measurements at unobserved locations can no longer be predicted well for the Markov-based policy since, unlike policy π_G, it is not capable of exploiting the (possibly large) horizontal correlation to improve prediction due to the Markov assumption. As a result, larger ERRs are incurred.

For the case of 3 robots (Fig. 9.8), the observations are as follows:

1. When the horizontal length-scale is kept constant (i.e., either at $\ell_x = 5$ m or at $\ell_x = 40.45$ m), decreasing the vertical length-scale from $\ell_y = 16$ m to $\ell_y = 5$ m (i.e., either from field 2 to field 1 or from field 4 to field 3) reduces the difference in ENT performance between the DMDP+FR-based policy $\widehat{\pi}$ and the non-Markovian

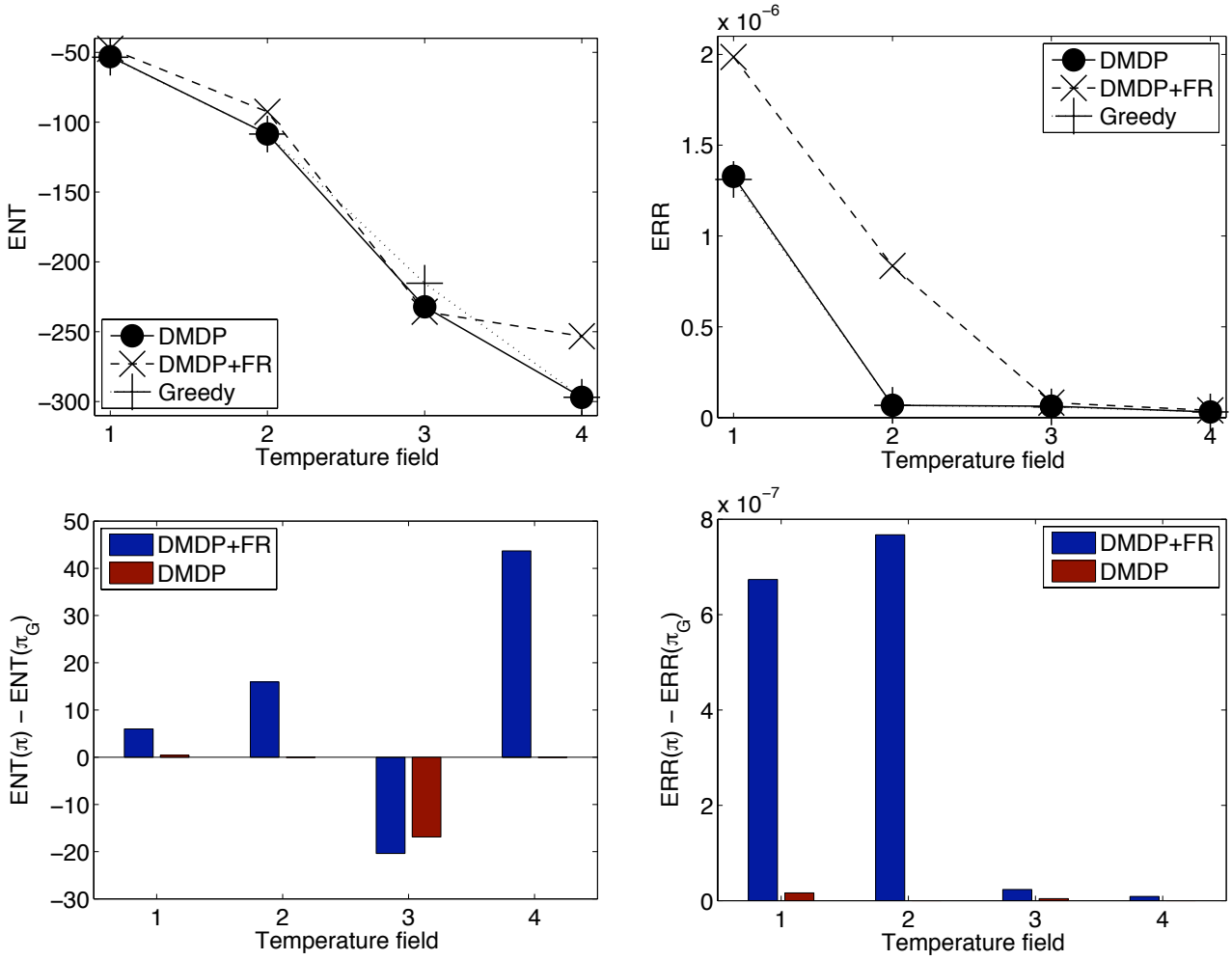


Figure 9.8: Performance comparison of DMDP-based, DMDP+FR-based, and non-Markovian greedy policies for the case of 3 robots.

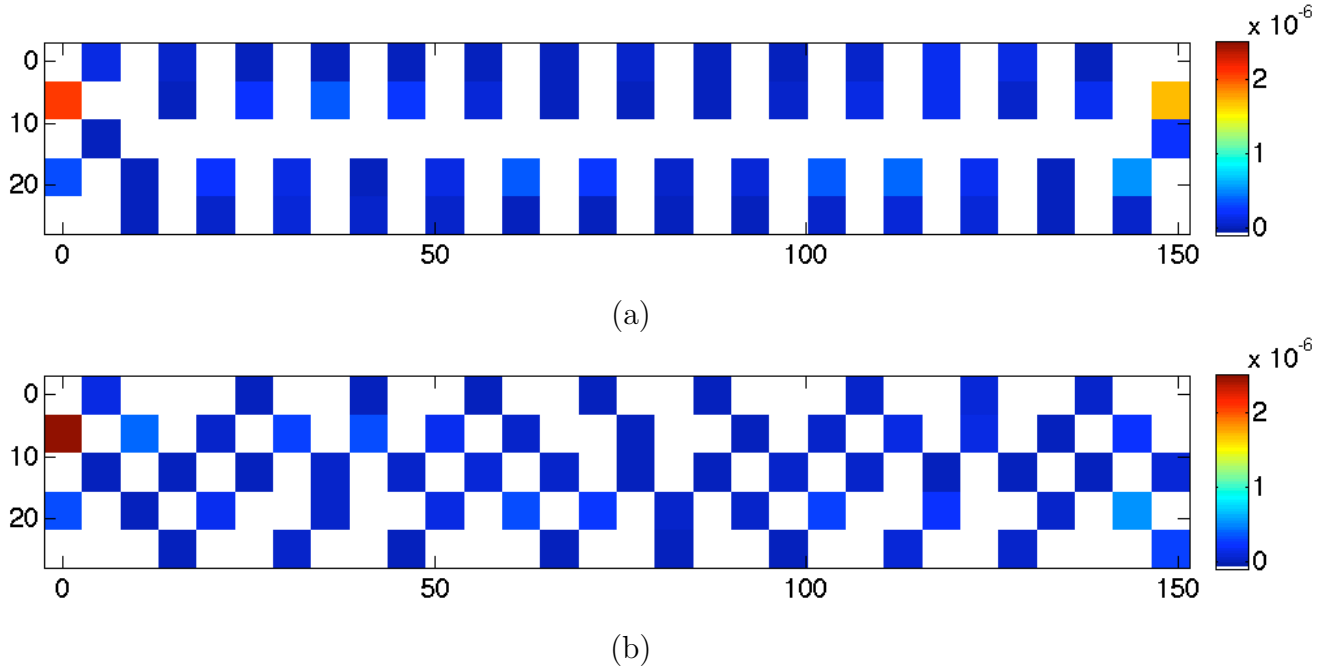


Figure 9.9: Squared relative error maps for field 3 with 3 robots showing white grid cells sampled by the (a) DMDP-based policy $\widetilde{\pi}$, and (b) non-Markovian greedy policy π_G.

greedy policy π_G: with small vertical correlation, the policy π_G loses its advantage of being able to exploit a large vertical correlation.

2. For field 3 with large horizontal length-scale ℓ_x and small vertical length-scale ℓ_y, a Markov-based policy achieves lower ENT than the non-Markovian greedy policy π_G: the unobserved locations are predominantly in groups of two for the DMDP-based policy $\widetilde{\pi}$ (Fig. 9.9a) and the DMDP+FR-based policy $\widehat{\pi}$, but the unobserved locations are mostly isolated for policy π_G. Consequently, the prior joint entropy of the measurements at unobserved locations is smaller for the Markov-based policy than for the non-Markovian greedy policy π_G. Since the horizontal correlation is large, the entropy reduction due to the sampled locations is relatively large for both policies. The resulting posterior map entropy is therefore lower for the Markov-based policy.

3. When the vertical length-scale is kept constant (i.e., either at $\ell_y = 5$ m or at $\ell_y = 16$ m), decreasing the horizontal length-scale from $\ell_x = 40.45$ m to $\ell_x = 5$ m (i.e., either from field 3 to field 1 or from field 4 to field 2) increases the difference in ERR performance between the DMDP+FR-based policy $\widehat{\pi}$ and the non-Markovian greedy policy π_G: for field 2, the DMDP+FR-based policy $\widehat{\pi}$ is not capable of exploiting the large vertical correlation to spread out the sampled locations in each column due to the conditional independence assumption (Section 9.4), thus incurring a higher ERR. In contrast, the DMDP-based policy $\widetilde{\pi}$ and the non-Markovian greedy policy π_G are capable of doing so. For field 1 with small horizontal and vertical length-scales, since a team of 3 robots can cover 60% of the exploration region, a uniform coverage achieved by policies $\widetilde{\pi}$ and π_G allows any unobserved location to be surrounded by sampled locations, thus offering better prediction and lower ERR. Due to the conditional independence assumption, the policy $\widehat{\pi}$ is not capable of uniform coverage.

4. The DMDP-based policy $\widetilde{\pi}$ can achieve ENT and ERR comparable to (if not, better

than) that of the non-Markovian greedy policy π_G for all temperature fields (e.g., Fig. 9.9).

Fig. 9.10 shows the time taken to derive the tested policies for exploring temperature field 4 with varying robot team sizes. For the other modified fields, the incurred time profiles are similar to that of Fig. 9.10. It can be observed that the time taken to derive the non-Markovian greedy policy π_G is longer than that needed to derive a Markov-based policy by more than an order of magnitude. Furthermore, note that Fig. 9.10 reports the average time taken to derive policy π_G over all possible starting robot locations. So, if the starting robot locations are unknown, the incurred time to derive policy π_G has to be increased by 5-, 10-, and 10-fold (i.e., ${}^5\mathrm{C}_k$-fold) for the case of 1, 2, and 3 robots, respectively. In contrast, the Markov-based policies cater to all possible starting robot locations. Therefore, the incurred time to derive a Markov-based policy remains unchanged even if the starting robot locations are unknown.

9.5.3 Summary of test results

The DMDP+FR-based policy $\widehat{\pi}$ can achieve very good ENT performance for temperature field 1 (i.e., of small horizontal and vertical length-scales) with any number of robots. This is in agreement with the result of Theorem 9.4.3. However, it achieves inferior ERR performance for the same field 1 because it is not capable of performing uniform coverage. It can achieve very good ERR performance for temperature field 4 with any number of robots by depending on the large horizontal and vertical correlations to predict well.

Though the DMDP-based policy $\widetilde{\pi}$ is not expected to achieve good ENT performance for a temperature field of large horizontal length-scale, it can achieve ENT performance comparable to that of the non-Markovian greedy policy π_G for all fields with any number of robots except for field 3 (i.e., of large horizontal length-scale and small vertical length-scale) with 1 robot. It can achieve very good ERR performance for fields 2 and 4 (i.e., of

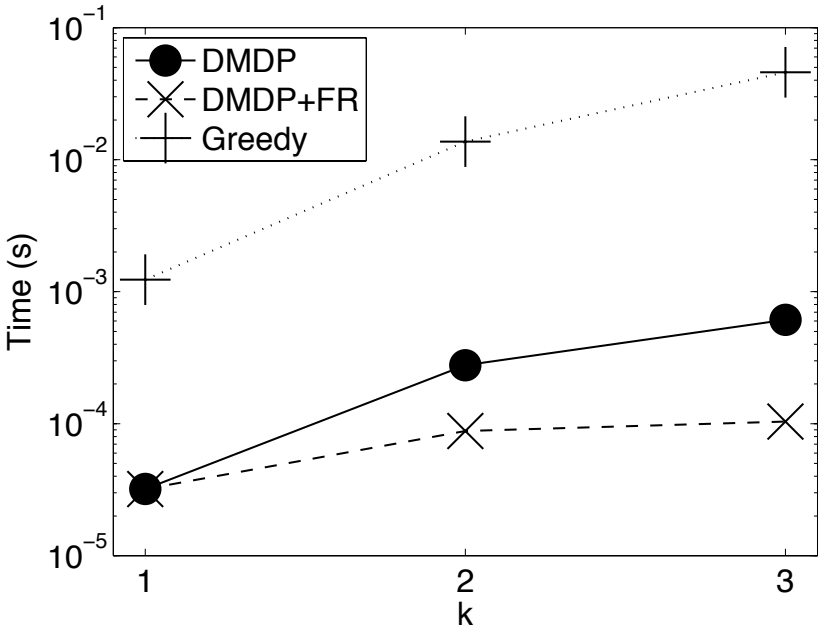


Figure 9.10: Graph of time taken to derive policy vs. number of robots k for temperature field 4.

large vertical length-scales) with 1 and 2 robots whereby it is capable of exploiting the large vertical correlation. It can also achieve very good ERR performance for all fields with 3 robots. As compared to the non-Markovian greedy policy π_G, it is faster to derive a Markov-based policy by more than an order of magnitude. Hence, a Markov-based path planner is more time-efficient for *in situ* real-time, high-resolution active sampling.

Chapter 10

Conclusion and Future Work

In this thesis, we have studied the following question:

> How does a robot team plan resource-constrained observation paths to minimize the map uncertainty of a hotspot field?

10.1 Summary of Contributions

To address the above question, the work in this thesis has provided the following novel contributions:

1. Formalization of MASP (Low *et al.*, 2008). MASP formalizes the exploration problem in a sequential decision-theoretic planning under uncertainty framework, which allows the performance of induced exploration policies of varying adaptivity to be theoretically analyzed and the performance advantage of a more adaptive policy to be realized. Through MASP, it is demonstrated that a more adaptive strategy can exploit clustering phenomena in a hotspot field to produce lower expected map uncertainty. To optimize the mean-squared error criterion, a MASP-based exploration strategy plans non-myopic adaptive observation paths that minimize the expected posterior map error or equivalently, maximize the expected map error reduction.

2. Formalization of iMASP (Low *et al.*, 2009). To optimize the entropy criterion, an iMASP-based exploration strategy plans non-myopic adaptive observation paths that minimize the expected posterior map entropy or equivalently, maximize the expected entropy of observation paths. Unlike MASP, the time complexity of solving the reward-maximizing iMASP approximately is independent of the map resolution.

3. Exploration strategies for learning hotspot field maps (Low *et al.*, 2008, 2009). The reward-maximizing MASP and iMASP allow observation selection properties of the induced exploration policies to be realized for sampling GP and ℓGP. These properties include adaptivity, hotspot sampling, and wide-area coverage.

4. Approximately optimal exploration strategies with performance guarantees (Low *et al.*, 2008, 2009). To handle continuous states, the convexity of reward-maximizing MASP and iMASP can be exploited to derive, in a computationally tractable manner, approximately optimal exploration policies with theoretical performance guarantees. Anytime algorithms based on approximate MASP and iMASP are then proposed to alleviate the computational difficulty that arises from their non-Markovian structure.

5. Quantifying "hotspotness". A "hotspotness" index is defined using the spatial correlation properties of the hotspot field. Consequently, this index can be related to the intensity, size, and diffuseness of the hotspots in the field.

6. Effects of spatial correlation on performance advantage of adaptivity. We have derived sufficient and necessary conditions of the spatial correlation properties of the hotspot field for adaptive exploration to yield no performance advantage.

7. Exploiting small spatial correlation with fast Markov-based exploration strategies. We have developed computationally efficient approximately optimal exploration strategies for sampling the GP by assuming the Markov property in iMASP planning. We provide

theoretical performance guarantees for the Markov-based policies, which improve with decreasing spatial correlation.

Note that the use of the exploration strategies described in contributions 1 through 4 are not limited to multi-robot teams; they can also be used for static sensor placements. The action space then becomes the set of different possible choices of placing the next sensor or group of sensors on the grid, which is often larger than the set of robot joint actions considered in this thesis.

10.2 Future Work

This section proposes a few directions that can be pursued as continuation to the work in this thesis.

1. Generalizing MASP to handle other optimizing criterion. We will identify other optimizing criterion with a resulting problem structure that can be exploited to derive computationally efficient or tractable exploration strategies.

2. Effects of spatial correlation on performance advantage of multi-stage adaptive exploration. We will derive necessary and sufficient conditions of the spatial correlation properties for multi-stage MASP and iMASP to yield no performance advantage. We will also develop theoretical bounds on the performance advantage of adaptivity that depend on the spatial correlation properties (i.e., length-scale, signal variance, and noise variance hyperparameters).

3. Exploiting small spatial correlation with fast Markov-based adaptive exploration strategies. We will devise computationally efficient Markov-based strategies for adaptive sampling of log-Gaussian process.

4. Improving map learning using auxiliary information. In our current work, the exploration strategies for learning the hotspot field map are guided by observations of a single environmental variable. However, the primary environmental variable to be mapped is often associated with some highly correlated auxiliary variable(s), which may be more densely sampled (e.g., due to cheaper sampling cost), sampled together with the primary variable during exploration, or available prior to exploration (e.g., via remote sensing). In algal bloom, the plankton density/abundance depends on the ocean conditions such as temperature, salinity, and nutrients (Apple *et al.*, 2008). In precision agriculture, the soil nutrients can be correlated to the crop yield (Webster and Oliver, 2007). We will investigate whether an exploration strategy can improve its map learning by further exploiting the observations of the correlated auxiliary variables.

5. Framing MASP as a classification/labeling problem. We will examine the feasibility of deploying a multi-robot science team for adaptive exploration to *label* (rather than map) a hotspot field. That is, instead of reconstructing a hotspot field map of continuous measurements, we are interested in predicting a map of binary labels. To label the measurements, we can make use of a predefined threshold. For example, a location is labeled 1 if its corresponding measurement is greater than the predefined threshold, and is labeled 0 otherwise. This threshold can be set as the permissible limit for pollutant concentration in pollution monitoring, for salinity or alkalinity level in precision agriculture, or for red-tide density in algal bloom. It is important to identify the potential regions where this threshold is likely to be exceeded due to economic, environmental, or health implications discussed in (Webster and Oliver, 2007). In this case, the exploration objective becomes one of classifying the hotspots correctly rather than predicting the field accurately.

Bibliography

Alvarez, A., Garau, B., and Caiti, A. (2007). Combining networks of drifting profiling floats and gliders for adaptive sampling of the ocean. In *Proc. IEEE ICRA*, pages 157–162. 8, 17, 18, 19, 20, 21, 26, 37

Apostolopoulos, D. S., Wagner, M. D., Shamah, B. N., Pedersen, L., Shillcutt, K., and Whittaker, W. L. (2000). Technology and field demonstration of robotic search for antarctic meteorites. *Int. J. Robot. Res.*, **19**(11), 1015–1032. 1

Apple, J. K., Smith, E. M., and Boyd, T. J. (2008). Temperature, salinity, nutrients, and the covariation of bacterial production and chlorophyll-a in estuarine ecosystems. *J. Coastal Res.*, **25**(sp1), 59–75. 140

Barto, A., Bradtke, S., and Singh, S. (1995). Learning to act using real-time dynamic programming. *Artif. Intell.*, **72**(1-2), 81–138. 69, 70

Batalin, M. A., Rahimi, M., Yu, Y., Liu, D., Kansal, A., Sukhatme, G. S., Kaiser, W. J., Hansen, M., Pottie, G. J., Srivastava, M., and Estrin, D. (2004). Call and response: Experiments in sampling the environment. In *Proc. ACM SenSys'04*, pages 25–38. 1, 17, 18, 19, 21, 26

Bertsekas, D. P. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA. 60

Birge, J. R. and Wets, R. J.-B. (1986). Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse. *Math. Programming Study*, **27**, 54–102. 62, 63

Bonet, B. and Geffner, H. (2003a). Faster heuristic search algorithms for planning with uncertainty and full feedback. In *Proc. IJCAI-03*, pages 1233–1238. 69, 70, 71

Bonet, B. and Geffner, H. (2003b). Labeled RTDP: Improving the convergence of real-time dynamic programming. In *Proc. ICAPS-03*, pages 12–21. 69, 70, 71

Boyan, J. A. and Littman, M. L. (2001). Exact solutions to time-dependent MDPs. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 1026–1032, Cambridge, MA. MIT Press. 9, 60, 61

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge Univ. Press, New York. 163, 164, 165

Brockwell, A. E. and Kadane, J. B. (2003). A gridding method for Bayesian sequential decision problems. *J. Computational and Graphical Statistics*, **12**(3), 566–584. 62, 63

Burgard, W., Moors, M., Stachniss, C., and Schneider, F. E. (2005). Coordinated multi-robot exploration. *IEEE Trans. Robotics*, **21**(3), 376–386. 1

Casey, M. S. and Sen, S. (2005). The scenario generation algorithm for multistage stochastic linear programming. *Math. Oper. Res.*, **30**(3), 615–631. 62, 63

Castano, R., Anderson, R. C., Estlin, T., DeCoste, D., Fisher, F., Gaines, D., Mazzoni, D., and Judd, M. (2003). Rover traverse science for increased mission science return. In *Proc. IEEE Aerospace Conference*, pages 3629–3636. 1

Chang, H., Fu, A. Q., Le, N. D., and Zidek, J. V. (2007). Designing environmental monitoring networks to measure extremes. *Environ. Ecol. Stat.*, **14**(3), 301–321. 3

Choset, H. (2001). Coverage for robotics: A survey of recent results. *Ann. Math. Artif. Intell.*, **31**(1-4), 113–126. 1

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *J. Artif. Intell. Res.*, **4**, 129–145. 4

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, NY. 157, 158

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, NY, second edition. 7, 38

De Oliveira, V. and Ecker, M. D. (2002). Bayesian hot spot detection in the presence of a spatial trend: Application to total nitrogen concentration in chesapeake bay. *Environmetrics*, **13**(1), 85–101. 84

Dupačová, J., Consigli, G., and Wallace, S. W. (2000). Scenarios for multistage stochastic programs. *Ann. Oper. Res.*, **100**(1-4), 25–53. 62, 63

Edirisinghe, N. C. P. (1999). Bound-based approximations in multistage stochastic programming: Nonanticipativity aggregation. *Ann. Oper. Res.*, **85**(1), 103–127. 62, 63

Englund, E. J. and Heravi, N. (1994). Phased sampling for soil remediation. *Environ. Ecol. Stat.*, **1**(3), 247–263. 3

Estlin, T., Mann, T., Gray, A., Rapideau, G., Castano, R., Chein, S., and Mjolsness, E. (1999). An integrated system for multi-rover scientific exploration. In *Proc. AAAI-99*, pages 613–620. 1

Estlin, T., Gaines, D., Fisher, F., and Castano, R. (2005). Coordinating multiple rovers with interdependent science objectives. In *Proc. AAMAS-05*, pages 879–886. 1

Feng, Z., Dearden, R., Meuleau, N., and Washington, R. (2004). Dynamic programming for structured continuous Markov decision problems. In *Proc. UAI-04*, pages 154–161. 9, 60, 61

Fiorelli, E., Leonard, N. E., Bhatta, P., Paley, D., Bachmayer, R., and Fratantoni, D. M. (2006). Multi-AUV control and adaptive sampling in Monterey Bay. *IEEE J. Oceanic Engineering*, **31**(4), 935–948. 1

Formisano, V., Atreya, S., Encrenaz, T., Ignatiev, N., and Giuranna, M. (2004). Detection of methane in the atmosphere of Mars. *Science*, **306**(5702), 1758–1761. 1

Frauendorfer, K. (1996). Barycentric scenario trees in convex multistage stochastic programming. *Math. Programming*, **75**(2), 277–293. 62, 63

Frauendorfer, K. and Haarbrücker, G. (2003). Solving sequences of refined multistage stochastic linear programs. *Ann. Oper. Res.*, **124**(1-4), 133–163. 62, 63

Frauendorfer, K. and Schürle, M. (2000). Term structure models in multistage stochastic programming: Estimation and approximation. *Ann. Oper. Res.*, **100**(1-4), 189–209. 62, 63

Glass, B. and Briggs, G. (2003). Evaluation of human vs. teleoperated robotic performance in field geology tasks at a Mars analog site. In *Proc. 7th i-SAIRAS-03*. 1

Golub, G. H. and Van Loan, C.-F. (1996). *Matrix Computations*. Johns Hopkins Univ. Press, third edition. 175

Guestrin, C., Hauskrecht, M., and Kveton, B. (2004). Solving factored MDPs with continuous and discrete variables. In *Proc. UAI-04*, pages 235–242. 60, 62

Guestrin, C., Krause, A., and Singh, A. P. (2005). Near-optimal sensor placements in Gaussian processes. In *Proc. ICML-05*, pages 265–272. 3, 48, 49, 79

Hauskrecht, M. and Kveton, B. (2004). Linear program approximations for factored continuous-state Markov decision processes. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 895–902, Cambridge, MA. MIT Press. 60, 62

Hohn, M. E. (1998). *Geostatistics and Petroleum Geology*. Springer, second edition. 37

Huang, C. C., Ziemba, W. T., and Ben-Tal, A. (1977). Bounds on the expectation of a convex function of a random variable: With applications to stochastic programming. *Oper. Res.*, **25**(2), 315–325. 63, 64, 160

Korf, R. (1990). Real-time heuristic search. *Artif. Intell.*, **42**(2-3), 189–211. 79, 107

Krasnopolsky, V. A., Maillard, J. P., and Owen, T. C. (2004). Detection of methane in the Martian atmosphere: Evidence for life? *Icarus*, **172**(2), 537–547. 1

Kveton, B. and Hauskrecht, M. (2006). Solving factored MDPs with exponential-family transition models. In *Proc. ICAPS-06*, pages 114–120. 9, 61, 62

Kveton, B., Hauskrecht, M., and Guestrin, C. (2006). Solving factored MDPs with hybrid state and action variables. *J. Artif. Intell. Res.*, **27**, 153–201. 9, 61, 62

Leonard, N. E., Paley, D., Lekien, F., Sepulchre, R., Fratantoni, D. M., and Davis, R. (2007). Collective motion, sensor networks and ocean sampling. *Proc. IEEE*, **95**(1), 48–74. 1, 17, 18, 19, 21, 26

Li, L. and Littman, M. L. (2005). Lazy approximation for solving continuous finite-horizon MDPs. In *Proc. AAAI-05*, pages 1175–1180. 9, 60, 61

Long, E. R. and Wilson, C. J. (1997). On the identification of toxic hot spots using measures of the sediment quality triad. *Marine Pollution Bulletin*, **34**(6), 373–374. 3, 84

Low, K. H., Gordon, G. J., Dolan, J. M., and Khosla, P. (2007). Adaptive sampling for multi-robot wide-area exploration. In *Proc. IEEE ICRA*, pages 755–760. 1, 17, 18, 19, 21, 26

Low, K. H., Dolan, J. M., and Khosla, P. (2008). Adaptive multi-robot wide-area exploration and mapping. In *Proc. AAMAS-08*, pages 23–30. 17, 18, 26, 137, 138

Low, K. H., Dolan, J. M., and Khosla, P. (2009). Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proc. ICAPS-09*. 17, 18, 138

Marecki, J., Koenig, S., and Tambe, M. (2007). A fast analytical algorithm for solving Markov decision processes with real-valued resources. In *Proc. IJCAI-07*, pages 2536–2541. 9, 60, 61

McCartney, R. and Sun, H. (2000). Sampling and estimation by multiple robots. In *Proc. ICMAS-00*, pages 415–416. 17, 18, 19, 21

McMahan, H. B., Likhachev, M., and Gordon, G. J. (2005). Bounded real-time dynamic programming: RTDP with monotone upper bounds and performance guarantees. In *Proc. ICML-05*, pages 569–576. 69, 70, 73

Meliou, A., Krause, A., Guestrin, C., Kaiser, W., and Hellerstein, J. M. (2007). Nonmyopic informative path planning in spatio-temporal models. In *Proc. AAAI-07*, pages 602–607. 8, 17, 18, 19, 20, 21, 26, 37

Müller, P., Berry, D. A., Grieve, A. P., Smith, M., and Krams, M. (2007). Simulation-based sequential Bayesian design. *J. Statistical Planning and Inference*, **137**(10), 3140–3150. 62, 63

Pedersen, S. M., Fountas, S., Have, H., and Blackmore, B. S. (2006). Agricultural robots system analysis and economic feasibility. *Precision Agri.*, **7**(4), 295–308. 1

Popa, D. O. and Lewis, F. L. (2008). Algorithms for robotic deployment of WSN in adaptive sampling applications. In Y. Li, M. T. Thai, and W. Wu, editors, *Wireless Sensor Networks and Applications*, Signals and Communication Technology, pages 35–64. Springer. 17, 18, 19, 20, 21, 26

Popa, D. O., Sanderson, A. C., Komerska, R. J., Mupparapu, S. S., Blidberg, D. R., and Chappel, S. G. (2004). Adaptive sampling algorithms for multiple autonomous underwater vehicles. In *Proc. IEEE/OES AUV*, pages 108–118. 26

Popa, D. O., Mysorewala, M. F., and Lewis, F. L. (2006). EKF-based adaptive sampling with mobile robotic sensor nodes. In *Proc. IEEE/RSJ IROS*, pages 2451–2456. 17, 18, 19, 20, 21

Press, W. H., Teukolsky, S. A., and Vetterling, W. T. (2002). *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge Univ. Press, New York, second edition. 33

Rahimi, M., Pon, R., Kaiser, W. J., Sukhatme, G. S., Estrin, D., and Srivastava, M. (2004). Adaptive sampling for environmental robotics. In *Proc. IEEE ICRA*, pages 3536–3544. 1, 17, 18, 19, 21, 26

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. 38, 41, 54, 77

Roy, N. and McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. In *ICML-01*, pages 441–448. 4

Rust, J. (1997). Using randomization to break the curse of dimensionality. *Econometrica*, **65**, 487–516. 60

Shapiro, A. (2006). On complexity of multistage stochastic programs. *Oper. Res. Lett.*, **34**(1), 1–8. 62, 63

Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling. *J. Applied Stat.*, **14**(2), 165–170. 7, 48, 50, 77

Singh, A., Nowak, R., and Ramanathan, P. (2006). Active learning for adaptive mobile sensing networks. In *Proc. IPSN*, pages 60–68. 17, 18, 19, 20, 21, 26

Singh, A., Krause, A., Guestrin, C., Kaiser, W., and Batalin, M. (2007). Efficient planning of informative paths for multiple robots. In *Proc. IJCAI-07*, pages 2204–2211. 8, 17, 18, 19, 20, 21, 26, 37, 48, 49

Smith, T. and Simmons, R. (2006). Focused real-time dynamic programming for MDPs: Squeezing more out of a heuristic. In *Proc. AAAI-06*. 69, 70, 73

Ståhl, A., Ringvall, and Lämås, T. (2000). Guided transect sampling for assessing sparse populations. *Forest Science*, **46**(1), 108–115. 107

Stewart, G. W. and Sun, J.-G. (1990). *Matrix Perturbation Theory*. Academic Press. 175

Sutton, R. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA. 60

Thompson, D. R. and Wettergreen, D. (2008). Intelligent maps for autonomous kilometer-scale science survey. In *Proc. International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)*. 107, 108

Thompson, W. L. (2004). *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters*. Island Press, Washington, DC. 3

Webster, R. and Oliver, M. (2007). *Geostatistics for Environmental Scientists*. John Wiley & Sons, Inc., NY, second edition. 37, 38, 140

Zhang, B. and Sukhatme, G. S. (2007). Adaptive sampling for estimating a scalar field using a robotic boat and a sensor network. In *Proc. IEEE ICRA*, pages 3673–3680. 17, 18, 19, 20, 21, 26

Appendix A

Proofs

A.1 Theorem 3.3.1

A basic property of integration is needed:

$$\int f(\mathbf{z_x} \mid d) \min_{\mathbf{a}\in\mathcal{A}(\mathbf{x})} Q(\mathbf{a}, d, \mathbf{x}, \mathbf{z_x})\ \mathrm{d}\mathbf{z_x} \leq \min_{\mathbf{a}\in\mathcal{A}(\mathbf{x})} \int f(\mathbf{z_x} \mid d)\ Q(\mathbf{a}, d, \mathbf{x}, \mathbf{z_x})\ \mathrm{d}\mathbf{z_x}$$

for any function Q. Note that the optimal value functions of MASP(1) in (3.7) can be expanded into a series of alternating minimum and expectation. By applying the above property repeatedly on (3.7), it can be derived that the induced optimal value $V_0^{\pi^1}(d_0)$ from solving MASP(1) will be less than or equal to the induced optimal value $V_0^{\pi^n}(d_0)$ from solving MASP(n) (3.11). In particular, the optimal value decreases monotonically in adaptivity as

shown below:

$$
\begin{aligned}
& V_0^{\pi^1}(d_0) \\
= \; & \min_{\mathbf{a}_0} \int f(\mathbf{z}_{\mathbf{x}_1} \mid d_0)\; V_1^{\pi^1}(d_1)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_1} \\
= \; & \min_{\mathbf{a}_0} \int f(\mathbf{z}_{\mathbf{x}_1} \mid d_0)\; \left[\min_{\mathbf{a}_1} \int f(\mathbf{z}_{\mathbf{x}_2} \mid d_1)\; V_2^{\pi^1}(d_2)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_2} \right]\; \mathrm{d}\mathbf{z}_{\mathbf{x}_1} \\
\leq \; & \min_{\mathbf{a}_0, \mathbf{a}_1} \int f(\mathbf{z}_{\mathbf{x}_1} \mid d_0)\; \int f(\mathbf{z}_{\mathbf{x}_2} \mid d_1)\; V_2^{\pi^1}(d_2)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_2}\; \mathrm{d}\mathbf{z}_{\mathbf{x}_1} \\
= \; & \min_{\mathbf{a}_0, \mathbf{a}_1} \int f(\mathbf{z}_{\mathbf{x}_{1:2}} \mid d_0)\; V_2^{\pi^1}(d_2)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:2}} \\
& \ldots \\
\leq \; & \min_{\mathbf{a}_0, \mathbf{a}_1} \int f(\mathbf{z}_{\mathbf{x}_{1:2}} \mid d_0)\; \Big[\min_{\mathbf{a}_2, \mathbf{a}_3} \int f(\mathbf{z}_{\mathbf{x}_{3:4}} \mid d_2)\; \ldots \\
& \Big[\min_{\mathbf{a}_{n-2}, \mathbf{a}_{n-1}} \int f(\mathbf{z}_{\mathbf{x}_{n-1:n}} \mid d_{n-2})\; V_{n/2}^{\pi^2}(d_n)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_{n-1:n}} \Big]\; \ldots\; \mathrm{d}\mathbf{z}_{\mathbf{x}_{3:4}} \Big] \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:2}} \qquad \text{(A.1)} \\
= \; & V_0^{\pi^2}(d_0) \\
& \ldots \\
\leq \; & \min_{\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2} \int f(\mathbf{z}_{\mathbf{x}_{1:3}} \mid d_0)\; \Big[\min_{\mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5} \int f(\mathbf{z}_{\mathbf{x}_{4:6}} \mid d_3)\; \ldots \\
& \Big[\min_{\mathbf{a}_{n-3}, \mathbf{a}_{n-2}, \mathbf{a}_{n-1}} \int f(\mathbf{z}_{\mathbf{x}_{n-2:n}} \mid d_{n-3})\; V_{n/3}^{\pi^3}(d_n)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_{n-2:n}} \Big]\; \ldots\; \mathrm{d}\mathbf{z}_{\mathbf{x}_{4:6}} \Big] \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:3}} \\
= \; & V_0^{\pi^3}(d_0) \\
& \ldots \\
\leq \; & \min_{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}} \int f(\mathbf{z}_{\mathbf{x}_{1:n}} \mid d_0)\; V_1^{\pi^n}(d_n)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:n}} \\
= \; & V_0^{\pi^n}(d_0)
\end{aligned}
$$

with the assignments $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i, \mathbf{a}_i)$ for $i \geq 0$ and $\mathbf{x}_{j+1:k+1} \leftarrow \tau(\mathbf{x}_{j:k}, \mathbf{a}_{j:k})$ for $0 \leq j < k$. The expression under the last inequality in (A.1) corresponds to the optimal value function of MASP(n) (3.11). Note that the inequalities follow from the abovementioned property.

A.2 Theorem 3.4.1

Proof by induction on i that $V_i^{\pi^1}(d_i) = \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_i} - U_i^{\pi^1}(d_i)$ for $i = n-1, \ldots, 0$.

Base case ($i = n-1$):

$$
\begin{aligned}
& V_{n-1}^{\pi^1}(d_{n-1}) \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \int f(\mathbf{z}_{\mathbf{x}_n} \mid d_{n-1})\ V_n^{\pi^1}(d_n)\ \mathrm{d}\mathbf{z}_{\mathbf{x}_n} \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \int f(\mathbf{z}_{\mathbf{x}_n} \mid d_{n-1})\ \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_n}\ \mathrm{d}\mathbf{z}_{\mathbf{x}_n} \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \mathbb{E}[\sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_n} \mid d_{n-1}] \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \sum_{x \in \mathcal{X}} \mathbb{E}[\sigma^2_{Z_x|d_n} \mid d_{n-1}] \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_{n-1}} - \mathrm{var}[\mu_{Z_x|d_n} \mid d_{n-1}] \\
&= \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_{n-1}} - \max_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \sum_{x \in \mathcal{X}} \mathrm{var}[\mu_{Z_x|d_n} \mid d_{n-1}] \\
&= \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_{n-1}} - U_{n-1}^{\pi^1}(d_{n-1})
\end{aligned}
\tag{A.2}
$$

with the assignment $\mathbf{x}_n \leftarrow \tau(\mathbf{x}_{n-1}, \mathbf{a}_{n-1})$. The first and second equalities follow from (3.7). The fourth equality follows from linearity of expectation. The fifth equality is due to the variance decomposition formula. Hence, the base case is true.

Inductive case: Suppose that

$$
V_i^{\pi^1}(d_i) = \sum_{x \in \mathcal{X}} \sigma^2_{Z_x|d_i}\ - U_i^{\pi^1}(d_i) \tag{A.3}
$$

is true. We have to prove that $V_{i-1}^{\pi^1}(d_{i-1}) = \sum_{x\in\mathcal{X}} \sigma^2_{Z_x|d_{i-1}} - U_{i-1}^{\pi^1}(d_{i-1})$ is true.

$$
\begin{aligned}
& V_{i-1}^{\pi^1}(d_{i-1}) \\
&= \min_{\mathbf{a}_{i-1}\in\mathcal{A}(\mathbf{x}_{i-1})} \int f(\mathbf{z}_{\mathbf{x}_i} \mid d_{i-1})\; V_i^{\pi^1}(d_i)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_i} \\
&= \min_{\mathbf{a}_{i-1}\in\mathcal{A}(\mathbf{x}_{i-1})} \mathbb{E}[V_i^{\pi^1}(d_i) \mid d_{i-1}] \\
&= \min_{\mathbf{a}_{i-1}\in\mathcal{A}(\mathbf{x}_{i-1})} \sum_{x\in\mathcal{X}} \mathbb{E}[\sigma^2_{Z_x|d_i} \mid d_{i-1}] - \mathbb{E}[U_i^{\pi^1}(d_i) \mid d_{i-1}] \\
&= \min_{\mathbf{a}_{i-1}\in\mathcal{A}(\mathbf{x}_{i-1})} \sum_{x\in\mathcal{X}} \Big(\sigma^2_{Z_x|d_{i-1}} - \mathrm{var}[\mu_{Z_x|d_i} \mid d_{i-1}]\Big) - \mathbb{E}[U_i^{\pi^1}(d_i) \mid d_{i-1}] \\
&= \sum_{x\in\mathcal{X}} \sigma^2_{Z_x|d_{i-1}} - \max_{\mathbf{a}_{i-1}\in\mathcal{A}(\mathbf{x}_{i-1})} \left(\sum_{x\in\mathcal{X}} \mathrm{var}[\mu_{Z_x|d_i} \mid d_{i-1}] + \mathbb{E}[U_i^{\pi^1}(d_i) \mid d_{i-1}]\right) \\
&= \sum_{x\in\mathcal{X}} \sigma^2_{Z_x|d_{i-1}} - U_{i-1}^{\pi^1}(d_{i-1})\ .
\end{aligned}
\tag{A.4}
$$

The first equality follows from (3.7). The third equality follows from linearity of expectation and (A.3). The fourth equality is due to the variance decomposition formula. The last equality is due to (3.14). Hence, the inductive case is true. It is clear from (A.2) and (A.4) that the optimal adaptive policy π^1 corresponding to the optimal value $V_0^{\pi^1}(d_0)$ coincides with that associated with $U_0^{\pi^1}(d_0)$.

A.3 Lemma 3.5.1

Previously, we have shown in Section 3.5.1 that $R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = \sum_{x\in\mathcal{X}} \sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}$ is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$. Alternatively, we can prove that $R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = \sum_{x\in\mathcal{X}} \mathrm{var}[\mu_{Z_x|d_{i+1}} \mid d_i]$ (3.15) is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$, which we will do here.

Case 1 (x is a component of $\mathbf{x}_{i+1}$): Then, $\mu_{Z_x|d_{i+1}} = Z_x$ and

$$\text{var}[\mu_{Z_x|d_{i+1}} \mid d_i] = \text{var}[Z_x \mid d_i] = \sigma^2_{Z_x|d_i} \ , \tag{A.5}$$

which is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ (3.20).

Case 2 (x is not a component of $\mathbf{x}_{i+1}$ or $\mathbf{x}_{0:i}$): Then, using (3.19),

$$\mu_{Z_x|d_{i+1}} = \mu_{Z_x} + \Sigma_{x\mathbf{x}_{0:i+1}}\Sigma^{-1}_{\mathbf{x}_{0:i+1}\mathbf{x}_{0:i+1}} \left\{ (\mathbf{z}_{\mathbf{x}_{0:i}}, \mathbf{z}_{\mathbf{x}_{i+1}})^\top - \boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{0:i+1}}} \right\} \ . \tag{A.6}$$

Let $\mathbf{w}$ be the last k components of $\Sigma_{x\mathbf{x}_{0:i+1}}\Sigma^{-1}_{\mathbf{x}_{0:i+1}\mathbf{x}_{0:i+1}}$, and Γ be the covariance matrix of $\mathbf{Z}_{\mathbf{x}_{i+1}}$ conditioned on d_i, that is,

$$\Gamma = \Sigma_{\mathbf{x}_{i+1}\mathbf{x}_{i+1}} - \Sigma_{\mathbf{x}_{i+1}\mathbf{x}_{0:i}}\Sigma^{-1}_{\mathbf{x}_{0:i}\mathbf{x}_{0:i}}\Sigma_{\mathbf{x}_{0:i}\mathbf{x}_{i+1}} \ . \tag{A.7}$$

Since $\mathbf{w}$ and Γ are both independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$, then, using (A.6) and (A.7),

$$\begin{aligned} \text{var}[\mu_{Z_x|d_{i+1}} \mid d_i] = &\ \text{var}[\mathbf{w}\mathbf{Z}^\top_{\mathbf{x}_{i+1}} \mid d_i] \\ = &\ \mathbf{w}\text{var}[\mathbf{Z}^\top_{\mathbf{x}_{i+1}} \mid d_i]\mathbf{w}^\top \\ = &\ \mathbf{w}\Gamma\mathbf{w}^\top \end{aligned} \tag{A.8}$$

is also independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$.

Hence, from (3.15), $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for $i = 0, \ldots, n-1$.

A.4 Lemma 3.5.2

Case 1 (x is a component of $\mathbf{x}_{i+1}$): Then, $\mu_{Y_x|d_{i+1}} = Y_x$. Since $Z_x = \log Y_x$ is normal, $\mu_{Y_x|d_{i+1}}$ is lognormal. Then, it follows from (3.19), (3.23), and (3.24) that $\text{var}[\mu_{Y_x|d_{i+1}} \mid d_i] = \sigma^2_{Y_x|d_i}$

depends on previously sampled data d_i.

Case 2 (x is not a component of $\mathbf{x}_{i+1}$ or $\mathbf{x}_{0:i}$): Then,

$$\mu_{Y_x|d_{i+1}} = \exp\{\mu_{Z_x|d_{i+1}} + \sigma^2_{Z_x|d_{i+1}}/2\}$$

and

$$\log \mu_{Y_x|d_{i+1}} = \mu_{Z_x|d_{i+1}} + \sigma^2_{Z_x|d_{i+1}}/2 \ . \tag{A.9}$$

Since $\mu_{Z_x|d_{i+1}}$ is a linear combination of normal random variables (i.e., components of $\mathbf{z}_{\mathbf{x}_{i+1}}$), $\log \mu_{Y_x|d_{i+1}}$ is normal. So, $\mu_{Y_x|d_{i+1}}$ is lognormal. Then,

$$\begin{aligned}
&\mathrm{var}[\mu_{Y_x|d_{i+1}} \mid d_i] \\
&= \mathbb{E}[\mu_{Y_x|d_{i+1}} \mid d_i]^2(\exp\{\mathrm{var}[\log \mu_{Y_x|d_{i+1}} \mid d_i]\} - 1) \\
&= \mu^2_{Y_x|d_i}(\exp\{\mathrm{var}[\mu_{Z_x|d_{i+1}} + \sigma^2_{Z_x|d_{i+1}}/2 \mid d_i]\} - 1) \\
&= \mu^2_{Y_x|d_i}(\exp\{\mathrm{var}[\mu_{Z_x|d_{i+1}} \mid d_i]\} - 1) \\
&= \mu^2_{Y_x|d_i}(\exp\{\mathbf{w}\Gamma\mathbf{w}^\top\} - 1) \\
&= \mu^2_{Y_x|d_i}(\exp\{\sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}\} - 1) \ .
\end{aligned} \tag{A.10}$$

The first equality follows from the lognormal $\mu_{Y_x|d_{i+1}}$ (3.24). The second equality follows from iterated expectations and (A.9). The third equality results from $\sigma^2_{Z_x|d_{i+1}}$ being independent of $\mathbf{z}_{\mathbf{x}_{i+1}}$. The fourth equality is due to (A.8). The last equality is due to the variance decomposition formula. It follows from (3.19) and (3.23) that $\mu_{y_x|d_i}$ depends on previously sampled data d_i. So, $\mathrm{var}[\mu_{Y_x|d_{i+1}} \mid d_i]$ depends on d_i.

Hence, from (3.15), $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ depends on d_i for $i = 0, \ldots, n-1$.

A.5 Theorem 4.2.1

Proof by induction on i that $V_i^{\pi^1}(d_i) = \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}} \mid d_i] - U_i^{\pi^1}(d_i)$ for $i = n-1, \ldots, 0$.

Base case ($i = n-1$):

$$
\begin{aligned}
& V_{n-1}^{\pi^1}(d_{n-1}) \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \int f(\mathbf{z}_{\mathbf{x}_n} \mid d_{n-1})\ V_n^{\pi^1}(d_n)\ \mathrm{d}\mathbf{z}_{\mathbf{x}_n} \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \int f(\mathbf{z}_{\mathbf{x}_n} \mid d_{n-1})\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} \mid d_n]\ \mathrm{d}\mathbf{z}_{\mathbf{x}_n} \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \mathbb{H}[\mathbf{Z}_{\mathbf{x}_n}, \mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} \mid d_{n-1}] - \mathbb{H}[\mathbf{Z}_{\mathbf{x}_n} \mid d_{n-1}] \\
&= \min_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n-1}} \mid d_{n-1}] - \mathbb{H}[\mathbf{Z}_{\mathbf{x}_n} \mid d_{n-1}] \\
&= \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n-1}} \mid d_{n-1}] - \max_{\mathbf{a}_{n-1} \in \mathcal{A}(\mathbf{x}_{n-1})} \mathbb{H}[\mathbf{Z}_{\mathbf{x}_n} \mid d_{n-1}] \\
&= \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n-1}} \mid d_{n-1}] - U_{n-1}^{\pi^1}(d_{n-1})\ .
\end{aligned}
$$

The first and second equalities follow from (3.7). The third equality is due to the chain rule for entropy (Cover and Thomas, 1991). The last equality is due to (3.14). Hence, the base case is true.

Inductive case: Suppose that

$$
V_i^{\pi^1}(d_i) = \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}} \mid d_i] - U_i^{\pi^1}(d_i) \tag{A.11}
$$

is true. We have to prove that $V_{i-1}^{\pi^1}(d_{i-1}) = \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i-1}} \mid d_{i-1}] - U_{i-1}^{\pi^1}(d_{i-1})$ is true.

$$
\begin{aligned}
& V_{i-1}^{\pi^1}(d_{i-1}) \\
&= \min_{\mathbf{a}_{i-1} \in \mathcal{A}(\mathbf{x}_{i-1})} \int f(\mathbf{z}_{\mathbf{x}_i} \mid d_{i-1})\ V_i^{\pi^1}(d_i)\ \mathrm{d}\mathbf{z}_{\mathbf{x}_i} \\
&= \min_{\mathbf{a}_{i-1} \in \mathcal{A}(\mathbf{x}_{i-1})} \int f(\mathbf{z}_{\mathbf{x}_i} \mid d_{i-1}) \left(\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}} \mid d_i] - U_i^{\pi^1}(d_i) \right) \mathrm{d}\mathbf{z}_{\mathbf{x}_i} \\
&= \min_{\mathbf{a}_{i-1} \in \mathcal{A}(\mathbf{x}_{i-1})} \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i-1}} \mid d_{i-1}] - \mathbb{H}[\mathbf{Z}_{\mathbf{x}_i} \mid d_{i-1}] - \int f(\mathbf{z}_{\mathbf{x}_i} \mid d_{i-1})\ U_i^{\pi^1}(d_i)\ \mathrm{d}\mathbf{z}_{\mathbf{x}_i} \\
&= \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i-1}} \mid d_{i-1}] - \max_{\mathbf{a}_{i-1} \in \mathcal{A}(\mathbf{x}_{i-1})} \left(\mathbb{H}[\mathbf{Z}_{\mathbf{x}_i} \mid d_{i-1}] \ + \int f(\mathbf{z}_{\mathbf{x}_i} \mid d_{i-1})\ U_i^{\pi^1}(d_i)\ \mathrm{d}\mathbf{z}_{\mathbf{x}_i} \right) \\
&= \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i-1}} \mid d_{i-1}] - U_{i-1}^{\pi^1}(d_{i-1})\ .
\end{aligned}
$$

The first equality follows from (3.7). The second equality follows from (A.11). The third equality follows from linearity of expectation and the chain rule for entropy (Cover and Thomas, 1991). The last equality is due to (3.14). Hence, the inductive case is true.

It is clear from above that the induced optimal adaptive policies from solving the cost-minimizing and reward-maximizing iMASP(1)'s coincide.

A.6 Equation 4.8

Since

$$
f(\mathbf{Z}_{\mathbf{x}_{i+1}} = \mathbf{z}_{\mathbf{x}_{i+1}} \mid d_i) = \frac{\exp\left\{ -\frac{1}{2} (\mathbf{z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i}) \Sigma^{-1}_{\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i} (\mathbf{z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i})^\top \right\}}{\sqrt{(2\pi)^k \ \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i} \mid}}\ ,
$$

$$
\begin{aligned}
&\mathbb{H}[\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i] \\
&= \mathbb{E}[-\log f(\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i) \mid d_i] \\
&= \mathbb{E}[\log \sqrt{(2\pi)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \frac{1}{2}(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i})\Sigma^{-1}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i})^\top \mid d_i] \\
&= \log \sqrt{(2\pi)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \frac{1}{2}\mathbb{E}[(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i})\Sigma^{-1}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i})^\top \mid d_i] \\
&= \log \sqrt{(2\pi)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \frac{1}{2}\mathbb{E}[\mathrm{tr}(\Sigma^{-1}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i})(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i})) \mid d_i] \\
&= \log \sqrt{(2\pi)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \frac{1}{2}\mathrm{tr}(\Sigma^{-1}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}\mathbb{E}[(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i})(\mathbf{Z}_{\mathbf{x}_{i+1}} - \mu_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}) \mid d_i]) \\
&= \log \sqrt{(2\pi)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \frac{1}{2}\mathrm{tr}(\Sigma^{-1}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}\Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}) \\
&= \log \sqrt{(2\pi)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \frac{1}{2}\mathrm{tr}(\mathbf{I}) \\
&= \log \sqrt{(2\pi)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \frac{k}{2} \\
&= \log \sqrt{(2\pi e)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid}\ .
\end{aligned}
$$

The fourth equality is due to the trace property $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

A.7 Equation 4.10

Using the Jacobian method of variable transformation,

$$
\begin{aligned}
f(\mathbf{Y}_{\mathbf{x}_{i+1}} \mid d_i) = &\quad f(\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i) \prod_{x \in \mathcal{X}'} \frac{dZ_x}{dY_x} \\
= &\quad f(\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i) \prod_{x \in \mathcal{X}'} \frac{1}{Y_x}
\end{aligned}
$$

where $\mathcal{X}' = \{x \mid x \text{ is a location component in } \mathbf{x}_{i+1}\}$. So,

$$\begin{aligned}
&\mathbb{H}[\mathbf{Y}_{\mathbf{x}_{i+1}} \mid d_i] \\
&= \mathbb{E}[-\log f(\mathbf{Y}_{\mathbf{x}_{i+1}} \mid d_i) \mid d_i] \\
&= \mathbb{E}[-\log \left(f(\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i) \prod_{x \in \mathcal{X}'} \frac{1}{Y_x} \right) \mid d_i] \\
&= \mathbb{E}[-\log f(\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i) + \sum_{x \in \mathcal{X}'} \log Y_x \mid d_i] \\
&= \mathbb{E}[-\log f(\mathbf{Z}_{\mathbf{x}_{i+1}} \mid d_i) \mid d_i] + \sum_{x \in \mathcal{X}'} \mathbb{E}[Z_x \mid d_i] \\
&= \log \sqrt{(2\pi e)^k \mid \Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mid} + \boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mathbf{1}^\top \ .
\end{aligned}$$

The fourth equality is due to the transformation $Z_x = \log Y_x$ and linearity of expectation. The fifth equality follows from (4.8).

A.8 Generalized Jensen and Edmundson-Madansky bounds

Theorem A.8.1 (Huang *et al.* (1977)). Let $W(\xi)$ be a convex function of ξ with the support $[a, b]$ that is subdivided at arbitrary points $b_0, \ldots, b_\nu$ (i.e., $a := b_0 < b_1 < \ldots < b_\nu =: b$). Let J_ν and M_ν denote the ν-fold generalized Jensen and Edmundson-Madansky bounds respectively:

$$J_\nu \triangleq \sum_{j=1}^{\nu} \alpha_j W(\beta_j), \quad M_\nu \triangleq \sum_{j=0}^{\nu} \delta_j W(b_j), \quad \nu = 1, 2, \ldots, \tag{A.12}$$

where

$$\alpha_j \triangleq \int_{b_{j-1}}^{b_j} f(\xi) d\xi, \quad \beta_j \triangleq \frac{1}{\alpha_j} \int_{b_{j-1}}^{b_j} \xi f(\xi) d\xi, \quad j = 1, \ldots, \nu$$

$$\delta_j \triangleq \alpha_j \left(\frac{\beta_j - b_{j-1}}{b_j - b_{j-1}} \right) + \alpha_{j+1} \left(\frac{b_{j+1} - \beta_{j+1}}{b_{j+1} - b_j} \right), j = 0, \ldots, \nu$$

and $\alpha_0 := \alpha_{\nu+1} := \beta_0 := \beta_{\nu+1} := b_{-1} := 0$. If the partition corresponding to $k+1$ is at least as fine as that corresponding to k for $k = 1, \ldots, \nu-1$, $J_1 \leq \ldots \leq J_\nu \leq \mathbb{E}[W(\xi)] \leq M_\nu \leq \ldots \leq M_1$.

The objective of this section is to construct piecewise-linear functions for bounding the convex function $W(\xi)$. Note that the expectation of $W(\xi)$ can be expressed as the sum of conditional expectations weighted on the intervals $[b_0, b_1], \ldots, [b_{\nu-1}, b_\nu]$ of the support $[b_0, b_\nu]$:

$$\mathbb{E}[W(\xi)] = \sum_{j=1}^{\nu} \alpha_j \, \mathbb{E}[W(\xi) \mid \xi \in [b_{j-1}, b_j] \,] \; .$$

For each interval $[b_{j-1}, b_j]$, a linear function $\underline{W}_j(\xi)$ can be constructed to lower-bound the convex function $W(\xi)$ such that it is tangential to $W(\xi)$ at some point β_j. So, its gradient has to be $W'(\beta_j)$. This gives the linear function

$$\underline{W}_j(\xi) = W(\beta_j) + W'(\beta_j)\{\xi - \beta_j\} \; .$$

Hence, we can lower-bound $\mathbb{E}[W(\xi) \mid \xi \in [b_{j-1}, b_j] \,]$ by the conditional expectation of $\underline{W}_j(\xi)$ on the interval $[b_{j-1}, b_j]$:

$$\begin{aligned} \mathbb{E}[\underline{W}_j(\xi) \mid \xi \in [b_{j-1}, b_j] \,] = \; & W(\beta_j) + W'(\beta_j)\{\mathbb{E}[\xi \mid \xi \in [b_{j-1}, b_j] \,] - \beta_j\} \\ = \; & \underline{W}_j(\mathbb{E}[\xi \mid \xi \in [b_{j-1}, b_j] \,]) \; . \end{aligned}$$

The largest lower bound can be obtained by differentiating $\underline{W}_j(\mathbb{E}[\xi \mid \xi \in [b_{j-1}, b_j] \,])$ with respect to β_j and setting it to 0. This gives

$$\beta_j = \mathbb{E}[\xi \mid \xi \in [b_{j-1}, b_j] \,] = \frac{1}{\alpha_j} \int_{b_{j-1}}^{b_j} \xi f(\xi) d\xi \; .$$

So, the ν-fold generalized Jensen bound can be derived as follows:

$$\begin{aligned}
\mathbb{E}[W(\xi)] = & \sum_{j=1}^{\nu} \alpha_j \ \mathbb{E}[W(\xi) \mid \xi \in [b_{j-1}, b_j] \ ] \\
\geq & \sum_{j=1}^{\nu} \alpha_j \ \mathbb{E}[\underline{W}_j(\xi) \mid \xi \in [b_{j-1}, b_j] \ ] \\
= & \sum_{j=1}^{\nu} \alpha_j \ W(\beta_j) \\
= & \ J_\nu \ .
\end{aligned}$$

For each interval $[b_{j-1}, b_j]$, a linear function $\overline{W}_j(\xi)$ can also be constructed to upper-bound the convex function $W(\xi)$ such that it crosses the two extreme points $(b_{j-1}, W(b_{j-1}))$ and $(b_j, W(b_j))$. Then, the linear function is of the form

$$\overline{W}_j(\xi) = \frac{W(b_j) - W(b_{j-1})}{b_j - b_{j-1}} \xi + \frac{b_j}{b_j - b_{j-1}} W(b_{j-1}) - \frac{b_{j-1}}{b_j - b_{j-1}} W(b_j) \ .$$

We can upper-bound $\mathbb{E}[W(\xi) \mid \xi \in [b_{j-1}, b_j] \ ]$ by the conditional expectation of $\overline{W}_j(\xi)$ on the interval $[b_{j-1}, b_j]$:

$$\begin{aligned}
\mathbb{E}[\overline{W}_j(\xi) \mid \xi \in [b_{j-1}, b_j] \ ] = & \ \frac{W(b_j) - W(b_{j-1})}{b_j - b_{j-1}} \mathbb{E}[\xi \mid \xi \in [b_{j-1}, b_j] \ ] + \\
& \ \frac{b_j}{b_j - b_{j-1}} W(b_{j-1}) - \frac{b_{j-1}}{b_j - b_{j-1}} W(b_j) \\
= & \ \frac{b_j - \beta_j}{b_j - b_{j-1}} W(b_{j-1}) + \frac{\beta_j - b_{j-1}}{b_j - b_{j-1}} W(b_j) \\
= & \ \overline{W}_j(\beta_j) \ .
\end{aligned}$$

So, the ν-fold generalized Edmundsen-Madansky bound can be derived as follows:

$$\begin{aligned}
\mathbb{E}[W(\xi)] &= \sum_{j=1}^{\nu} \alpha_j \ \mathbb{E}[W(\xi) \mid \xi \in [b_{j-1}, b_j] \ ] \\
&\leq \sum_{j=1}^{\nu} \alpha_j \ \mathbb{E}[\overline{W}_j(\xi) \mid \xi \in [b_{j-1}, b_j] \ ] \\
&= \sum_{j=1}^{\nu} \alpha_j \ \left(\frac{b_j - \beta_j}{b_j - b_{j-1}} W(b_{j-1}) + \frac{\beta_j - b_{j-1}}{b_j - b_{j-1}} W(b_j) \right) \\
&= \sum_{j=0}^{\nu} \delta_j W(b_j) \\
&= M_\nu \ .
\end{aligned}$$

A.9 Lemma 5.3.1

We first show that $R(x_{i+1}, d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ for $i = 0, \ldots, t$.

For MASP($\frac{1}{k}$) (5.1), we know from (3.25) that

$$R(x_{i+1}, d_i) = \sum_{x \in \mathcal{X}} \mu^2_{Y_x|d_i} (\exp\{\sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}\} - 1) \ .$$

From (3.19) and (3.23), $\mu^2_{Y_x|d_i}$ is the exponential of an affine function of $\mathbf{z}_{\mathbf{x}_{0:i}}$. Hence, it is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ ((Boyd and Vandenberghe, 2004), pp. 79). From (3.20), the posterior variances $\sigma^2_{Z_x|d_i}$ and $\sigma^2_{Z_x|d_{i+1}}$ are independent of $\mathbf{z}_{\mathbf{x}_{0:i}}$. So, $\exp\{\sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}\} - 1$ is a constant term. Since the sum with respect to x preserves convexity, $R(x_{i+1}, d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$.

For iMASP($\frac{1}{k}$) (5.1), we know from (4.7) that

$$R(x_{i+1}, d_i) \stackrel{\triangle}{=} \mathbb{H}[Y_{x_{i+1}}|d_i] = \log \sqrt{2\pi e \sigma^2_{Z_{x_{i+1}}|d_i}} + \mu_{Z_{x_{i+1}}|d_i} \ .$$

From (3.19), the posterior mean $\mu_{Z_{x_{i+1}}|d_i}$ is an affine function of $\mathbf{z}_{\mathbf{x}_{0:i}}$. Hence, it is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ ((Boyd and Vandenberghe, 2004), pp. 71). From (3.20), the posterior variance $\sigma^2_{Z_{x_{i+1}}|d_i}$ is independent of $\mathbf{z}_{\mathbf{x}_{0:i}}$. So, $\log\sqrt{2\pi e\sigma^2_{Z_{x_{i+1}}|d_i}}$ is a constant term. Therefore, $R(x_{i+1}, d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$.

We will revert to using $Z_{x_{i+1}}$ in MASP($\frac{1}{k}$) and iMASP($\frac{1}{k}$) (5.1) for ℓGP (i.e., by transforming $Z_{x_{i+1}} = \log Y_{x_{i+1}}$). The induction proof below holds for both MASP($\frac{1}{k}$) and iMASP($\frac{1}{k}$).

Proof by induction on i that $U_i(d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ for $i = t, \ldots, 0$.

Base case ($i = t$): As proven above, $R(x_{t+1}, d_t)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:t}}$. Then, the pointwise maximum of $R(x_{t+1}, d_t)$ (i.e., $\max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t)$) is convex in $\mathbf{z}_{\mathbf{x}_{0:t}}$ ((Boyd and Vandenberghe, 2004), pp. 81). Therefore, $U_t(d_t)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:t}}$. The base case is true.

Inductive case: Suppose that $U_{i+1}(d_{i+1})$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i+1}}$. We have to prove that $U_i(d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$.

From (5.1), the expectation under the normal variable $Z_{x_{i+1}}$ with posterior mean $\mu_{Z_{x_{i+1}}|d_i}$ and variance $\sigma^2_{Z_{x_{i+1}}|d_i}$ can be expressed in terms of the standard normal variable $Z = (Z_{x_{i+1}} - \mu_{Z_{x_{i+1}}|d_i})/\sigma^2_{Z_{x_{i+1}}|d_i}$:

$$\int f(Z_{x_{i+1}} = z_{x_{i+1}}|d_i)\, U_{i+1}(d_i, x_{i+1}, z_{x_{i+1}})\, \mathrm{d}z_{x_{i+1}} = \int f(z)\, U_{i+1}(d_i, x_{i+1}, \mu_{Z_{x_{i+1}}|d_i} + \sigma^2_{Z_{x_{i+1}}|d_i} z)\, \mathrm{d}z\,.$$

Since d_i and $\mu_{Z_{x_{i+1}}|d_i} + \sigma^2_{Z_{x_{i+1}}|d_i} z$ are affine in $\mathbf{z}_{\mathbf{x}_{0:i}}$ and $U_{i+1}(d_{i+1})$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i+1}}$ by assumption, $U_{i+1}(d_i, x_{i+1}, \mu_{Z_{x_{i+1}}|d_i} + \sigma^2_{Z_{x_{i+1}}|d_i} z)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ because vector composition operation preserves convexity[1] ((Boyd and Vandenberghe, 2004), pp. 86). Since

[1]Note that $U_{i+1}(d_{i+1})$ does not have to be non-decreasing in each argument because d_i and $\mu_{Z_{x_{i+1}}|d_i} + \sigma^2_{Z_{x_{i+1}}|d_i} z$ are affine in $\mathbf{z}_{\mathbf{x}_{0:i}}$.

$U_{i+1}(d_i, x_{i+1}, \mu_{Z_{x_{i+1}}|d_i} + \sigma^2_{Z_{x_{i+1}}|d_i} z)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ for each z, $\int f(z)\, U_{i+1}(d_i, x_{i+1}, \mu_{Z_{x_{i+1}}|d_i} + \sigma^2_{Z_{x_{i+1}}|d_i} z)\, \mathrm{d}z$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ because integration preserves convexity ((Boyd and Vandenberghe, 2004), pp. 79). So, $\int f(z_{x_{i+1}}|d_i)\, U_{i+1}(d_i, x_{i+1}, z_{x_{i+1}})\, \mathrm{d}z_{x_{i+1}}$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$. From above, $R(x_{i+1}, d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$. Then, the pointwise maximum of $R(x_{i+1}, d_i) + \int f(z_{x_{i+1}} \mid d_i)\, U_{i+1}(d_i, x_{i+1}, z_{x_{i+1}})\, \mathrm{d}z_{x_{i+1}}$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$. Therefore, $U_i(d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$. The inductive case is true.

A.10 Theorem 5.3.1

Proof by induction on i that $\underline{U}_i^{\nu}(d_i) \leq \underline{U}_i^{\nu+1}(d_i) \leq U_i(d_i)$ for $i = t, \ldots, 0$.

Base case ($i = t$): From (5.1) and (5.4), $\underline{U}_t^{\nu}(d_t) = \underline{U}_t^{\nu+1}(d_t) = U_t(d_t) = \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t)$. Hence, the base case is true.

Inductive case: Suppose that $\underline{U}_{i+1}^{\nu}(d_{i+1}) \leq \underline{U}_{i+1}^{\nu+1}(d_{i+1}) \leq U_{i+1}(d_{i+1})$ is true. We have to prove that $\underline{U}_i^{\nu}(d_i) \leq \underline{U}_i^{\nu+1}(d_i) \leq U_i(d_i)$ is true.

We will first show that $\underline{U}_i^{\nu+1}(d_i) \leq U_i(d_i)$.

$$
\begin{aligned}
\underline{U}_i^{\nu+1}(d_i) = & \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu+1} \underline{p}_{x_{i+1}}^{[j]} \underline{U}_{i+1}^{\nu+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]}) \\
\leq & \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu+1} \underline{p}_{x_{i+1}}^{[j]} U_{i+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]}) \\
\leq & \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \mathbb{E}[U_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}}) \mid d_i] \\
= & \; U_i(d_i) \; .
\end{aligned}
$$

The first equality is due to (5.4). The first inequality follows from assumption, that is, $\underline{U}_{i+1}^{\nu+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]}) \leq U_{i+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]})$. The second inequality follows from Lemma 5.3.1

that $U_{i+1}(d_i, x_{i+1}, z_{x_{i+1}})$ is convex in $z_{x_{i+1}}$ for ℓGP, and the generalized Jensen bound (5.2). The last equality is due to (5.1).

We will now prove that $\underline{U}_i^{\nu}(d_i) \leq \underline{U}_i^{\nu+1}(d_i)$.

$$\begin{aligned}
\underline{U}_i^{\nu}(d_i) = & \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}_{x_{i+1}}^{[j]} \underline{U}_{i+1}^{\nu}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]}) \\
\leq & \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}_{x_{i+1}}^{[j]} \underline{U}_{i+1}^{\nu+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]}) \\
\leq & \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{\ell=1}^{\nu+1} \underline{p}_{x_{i+1}}^{[\ell]} \underline{U}_{i+1}^{\nu+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[\ell]}) \\
= & \underline{U}_i^{\nu+1}(d_i) \ .
\end{aligned}$$

The equalities are due to (5.4). The first inequality follows from assumption, that is, $\underline{U}_{i+1}^{\nu}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]}) \leq \underline{U}_{i+1}^{\nu+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]}))$. We need the result that $\underline{U}_i^{\nu+1}(d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ for $i = 0, \ldots, t$ for the second inequality to hold. The proof[2] is similar to that of Lemma 5.3.1. Consequently, since $\underline{U}_{i+1}^{\nu+1}(d_i, x_{i+1}, z_{x_{i+1}})$ is convex in $z_{x_{i+1}}$ and $\mathcal{Z}_{x_{i+1}}^{\nu+1}$ is obtained by splitting one of the intervals in $\mathcal{Z}_{x_{i+1}}^{\nu}$, the second inequality results. The inductive case is thus true.

The proof of $U_i(d_i) \leq \overline{U}_i^{\nu+1}(d_i) \leq \overline{U}_i^{\nu}(d_i)$ for $i = t, \ldots, 0$ is similar to the above except that the inequalities are reversed.

A.11 Theorem 5.3.2

Proof by induction on i that $\underline{U}_i^{\nu}(d_i) \leq U_i^{\pi^{\frac{1}{k}}}(d_i) \leq U_i(d_i)$ for $i = t, \ldots, 0$.

Base case $(i = t)$: $U_t^{\pi^{\frac{1}{k}}}(d_t) = R(\tau(\mathbf{x}_t, \underline{\pi}_t^{\frac{1}{k}}(d_t)), d_t) = \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} R(x_{t+1}, d_t) = \underline{U}_t^{\nu}(d_t) = U_t(d_t)$.

[2] The lower approximate problem $\underline{\text{MASP}}(\frac{1}{k})/\underline{i\text{MASP}}(\frac{1}{k})$ and upper approximate problem $\overline{\text{MASP}}(\frac{1}{k})/\overline{i\text{MASP}}(\frac{1}{k})$ differ from $\text{MASP}(\frac{1}{k})/i\text{MASP}(\frac{1}{k})$ (5.1) by the non-negative weighted sum (instead of the expectation), which also preserves convexity.

Hence, the base case is true.

Inductive case: Suppose that $\underline{U}^{\nu}_{i+1}(d_{i+1}) \leq U^{\underline{\pi}^{\frac{1}{k}}}_{i+1}(d_{i+1}) \leq U_{i+1}(d_{i+1})$ is true. We have to prove that $\underline{U}^{\nu}_{i}(d_{i}) \leq U^{\underline{\pi}^{\frac{1}{k}}}_{i}(d_{i}) \leq U_{i}(d_{i})$ is true.

We will first show that $U^{\underline{\pi}^{\frac{1}{k}}}_{i}(d_i) \leq U_i(d_i)$.

$$\begin{aligned}
U^{\underline{\pi}^{\frac{1}{k}}}_{i}(d_i) = &\ R(\tau(\mathbf{x}_i, \underline{\pi}^{\frac{1}{k}}_i(d_i)), d_i) + \mathbb{E}[U^{\underline{\pi}^{\frac{1}{k}}}_{i+1}(d_{i+1}) \mid d_i] \\
\leq &\ R(\tau(\mathbf{x}_i, \underline{\pi}^{\frac{1}{k}}_i(d_i)), d_i) + \mathbb{E}[U_{i+1}(d_{i+1}) \mid d_i] \\
\leq &\ \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \mathbb{E}[U_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}}) \mid d_i] \\
= &\ U_i(d_i) \ .
\end{aligned}$$

The first inequality follows from assumption (i.e., $U^{\underline{\pi}^{\frac{1}{k}}}_{i+1}(d_{i+1}) \leq U_{i+1}(d_{i+1})$). We will now prove that $\underline{U}^{\nu}_{i}(d_{i}) \leq U^{\underline{\pi}^{\frac{1}{k}}}_{i}(d_{i})$. It requires the observation that $U^{\pi}_i(d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ for $i = 0, \ldots, t$, which can be shown in a similar manner as that of Lemma 5.3.1 without the max operator.

$$\begin{aligned}
\underline{U}^{\nu}_{i}(d_i) = &\ \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} \underline{U}^{\nu}_{i+1}(d_i, x_{i+1}, \underline{z}^{[j]}_{x_{i+1}}) \\
= &\ R(\tau(\mathbf{x}_i, \underline{\pi}^{\frac{1}{k}}_i(d_i)), d_i) + \sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} \underline{U}^{\nu}_{i+1}(d_i, \tau(\mathbf{x}_i, \underline{\pi}^{\frac{1}{k}}_i(d_i)), \underline{z}^{[j]}_{x_{i+1}}) \\
\leq &\ R(\tau(\mathbf{x}_i, \underline{\pi}^{\frac{1}{k}}_i(d_i)), d_i) + \sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} U^{\underline{\pi}^{\frac{1}{k}}}_{i+1}(d_i, \tau(\mathbf{x}_i, \underline{\pi}^{\frac{1}{k}}_i(d_i)), \underline{z}^{[j]}_{x_{i+1}}) \\
\leq &\ R(\tau(\mathbf{x}_i, \underline{\pi}^{\frac{1}{k}}_i(d_i)), d_i) + \mathbb{E}[U^{\underline{\pi}^{\frac{1}{k}}}_{i+1}(d_{i+1}) \mid d_i] \\
= &\ U^{\underline{\pi}^{\frac{1}{k}}}_{i}(d_i) \ .
\end{aligned}$$

The first inequality follows from assumption (i.e., $\underline{U}^{\nu}_{i+1}(d_{i+1}) \leq U^{\underline{\pi}^{\frac{1}{k}}}_{i+1}(d_{i+1})$). The second inequality follows from the observation that $U^{\underline{\pi}^{\frac{1}{k}}}_{i+1}(d_{i+1})$ is convex in $z_{x_{i+1}}$ and Theorem A.8.1. The inductive case is thus true.

A.12 Monotonicity of Lower Heuristic Bound

$$\begin{aligned}\underline{H}_i(d_i) = &\ R(x^*_{i+1}, d_i) + \underline{H}_{i+1}(d_i, x^*_{i+1}, \mathbb{E}[Z_{x^*_{i+1}} \mid d_i]) \\ \leq &\ \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \underline{H}_{i+1}(d_i, x_{i+1}, \mathbb{E}[Z_{x_{i+1}} \mid d_i]) \\ \leq &\ \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \mathbb{E}[\underline{H}_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}}) \mid d_i]\end{aligned}$$

for stage $i = 0, \ldots, t-1$. The second inequality follows from the convexity of $\underline{H}_{i+1}(d_i, x_{i+1}, z_{x_{i+1}})$ in $z_{x_{i+1}}$ and Theorem A.8.1 (i.e., Jensen bound). Consequently, by Theorem A.8.1,

$$\underline{H}_i(d_i) \leq \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}^{[j]}_{x_{i+1}} \underline{H}_{i+1}(d_i, x_{i+1}, \underline{z}^{[j]}_{x_{i+1}})$$

with $\underline{p}^{[j]}_{x_{i+1}}$ and $\underline{z}^{[j]}_{x_{i+1}}$ defined according to that of $\underline{\text{MASP}}(\frac{1}{k})$ (5.4).

A.13 Monotonicity of Upper Heuristic Bound

$$\begin{aligned}&\max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \mathbb{E}[\overline{H}_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}}) \mid d_i] \\ &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} \sum_{x \in \mathcal{X}} \text{var}[\mu_{Y_x|d_i, x_{i+1}, \exp\{Z_{x_{i+1}}\}} \mid d_i] + \mathbb{E}[\sum_{x \in \mathcal{X}} \sigma^2_{Y_x|d_i, x_{i+1}, \exp\{Z_{x_{i+1}}\}} \mid d_i] \\ &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} \sum_{x \in \mathcal{X}} \text{var}[\mu_{Y_x|d_i, x_{i+1}, Y_{x_{i+1}}} \mid d_i] + \mathbb{E}[\sum_{x \in \mathcal{X}} \sigma^2_{Y_x|d_i, x_{i+1}, Y_{x_{i+1}}} \mid d_i] \\ &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} \sum_{x \in \mathcal{X}} \text{var}[\mu_{Y_x|d_i, x_{i+1}, Y_{x_{i+1}}} \mid d_i] + \sum_{x \in \mathcal{X}} \mathbb{E}[\sigma^2_{Y_x|d_i, x_{i+1}, Y_{x_{i+1}}} \mid d_i] \\ &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} \sum_{x \in \mathcal{X}} \text{var}[\mu_{Y_x|d_i, x_{i+1}, Y_{x_{i+1}}} \mid d_i] + \sum_{x \in \mathcal{X}} (\sigma^2_{Y_x|d_i} - \text{var}[\mu_{Y_x|d_i, x_{i+1}, Y_{x_{i+1}}} \mid d_i]) \\ &= \sum_{x \in \mathcal{X}} \sigma^2_{Y_x|d_i} \\ &= \overline{H}_i(d_i)\end{aligned}$$

for stage $i = 0, \ldots, t-2$. Since $\overline{H}_{i+1}(d_i, x_{i+1}, z_{x_{i+1}})$ is convex in $z_{x_{i+1}}$, by Theorem A.8.1 (i.e., Jensen bound),

$$\overline{H}_i(d_i) \geq \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} R(x_{i+1}, d_i) + \sum_{j=1}^{\nu} \underline{p}_{x_{i+1}}^{[j]} \overline{H}_{i+1}(d_i, x_{i+1}, \underline{z}_{x_{i+1}}^{[j]})$$

with $\underline{p}_{x_{i+1}}^{[j]}$ and $\underline{z}_{x_{i+1}}^{[j]}$ defined according to that of $\underline{\text{MASP}}(\frac{1}{k})$ (5.4). When $i = t-1$,

$$\begin{aligned}
&\overline{H}_{t-1}(d_{t-1}) \\
&= \sum_{x \in \mathcal{X}} \sigma^2_{Y_x | d_{t-1}} \\
&\geq U_{t-1}(d_{t-1}) \\
&\geq \underline{U}^{\nu}_{t-1}(d_{t-1}) \\
&= \max_{\mathbf{a}_{t-1} \in \mathcal{A}'(\mathbf{x}_{t-1})} R(x_t, d_{t-1}) + \sum_{j=1}^{\nu} \underline{p}_{x_t}^{[j]} \ \underline{U}^{\nu}_t(d_{t-1}, x_t, \underline{z}_{x_t}^{[j]}) \\
&= \max_{\mathbf{a}_{t-1} \in \mathcal{A}'(\mathbf{x}_{t-1})} R(x_t, d_{t-1}) + \sum_{j=1}^{\nu} \underline{p}_{x_t}^{[j]} \ \overline{H}_t(d_{t-1}, x_t, \underline{z}_{x_t}^{[j]})
\end{aligned}$$

The first inequality follows from Theorem 3.4.1 and the second inequality is due to Theorem 5.3.1.

A.14 Theorem 8.1.1

For the squared exponential covariance function (7.1),

Case of $\ell = 0$:

$$\sigma_{Z_x Z_u} = \begin{cases} \sigma_s^2 + \sigma_n^2 & \text{if } x = u, \\ 0 & \text{otherwise.} \end{cases} \tag{A.13}$$

Case of $\ell = \infty$:

$$\sigma_{Z_x Z_u} = \begin{cases} \sigma_s^2 + \sigma_n^2 & \text{if } x = u, \\ \sigma_s^2 & \text{otherwise.} \end{cases} \tag{A.14}$$

A.14.1 MASP

Recall from Appendix A.4 that the reward function (3.15) is

$$R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = \sum_{x \in \mathcal{X}} \mu_{Y_x|d_i}^2 (\exp\{\sigma_{Z_x|d_i}^2 - \sigma_{Z_x|d_{i+1}}^2\} - 1) \ .$$

Case of $\ell = 0$: If x is a component of $\mathbf{x}_{0:i}$, $\sigma_{Z_x|d_i}^2 = \sigma_{Z_x|d_{i+1}}^2 = 0$. If x is not a component of $\mathbf{x}_{0:i+1}$, $\sigma_{Z_x|d_i}^2 = \sigma_{Z_x|d_{i+1}}^2 = \sigma_{Z_x Z_x}$. If x is a component of $\mathbf{x}_{i+1}$ but not a component of $\mathbf{x}_{0:i}$, $\sigma_{Z_x|d_i}^2 = \sigma_{Z_x Z_x}$ and $\sigma_{Z_x|d_{i+1}}^2 = 0$. It follows that $R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = \sum_{x \in \mathcal{X}'} \mu_{Y_x}^2 (\exp\{\sigma_{Z_x Z_x}\} - 1) = \sum_{x \in \mathcal{X}'} \sigma_{Y_x Y_x}$ where $\mathcal{X}' = \{x \mid x \text{ is a location component of } \mathbf{x}_{i+1} \text{ but not of } \mathbf{x}_{0:i}\}$. Hence, Lemma 3.5.1 is satisfied. As a result, Theorem 3.5.1 holds. Therefore, π^1 is non-adaptive. The reduction of MASP(1) to a single-staged MASP(n) is similar to that of (3.22).

Case of $\ell = \infty$: To simplify exposition, we assume that no prior data d_0 are available. If x is a component of $\mathbf{x}_{1:i}$, $\sigma_{Z_x|d_i}^2 = \sigma_{Z_x|d_{i+1}}^2 = 0$. If x is not a component of $\mathbf{x}_{1:i+1}$,

$$\begin{aligned} \sigma_{Z_x|d_i}^2 &= \sigma_{Z_x Z_x} - \Sigma_{x \mathbf{x}_{1:i}} \Sigma_{\mathbf{x}_{1:i} \mathbf{x}_{1:i}}^{-1} \Sigma_{\mathbf{x}_{1:i} x} \\ &= \sigma_n^2 \left(1 + \frac{\sigma_s^2}{ki\sigma_s^2 + \sigma_n^2}\right) \ . \end{aligned} \tag{A.15}$$

The second equality follows from the covariance vector $\Sigma_{x \mathbf{x}_{1:i}}$ with components σ_s^2, and the covariance matrix $\Sigma_{\mathbf{x}_{1:i} \mathbf{x}_{1:i}}$ with diagonal components $\sigma_s^2 + \sigma_n^2$ and off-diagonal components σ_s^2. As a result, $\Sigma_{\mathbf{x}_{1:i} \mathbf{x}_{1:i}}^{-1}$ has diagonal components $\dfrac{(ki-1)\sigma_s^2 + \sigma_n^2}{\sigma_n^2(ki\sigma_s^2 + \sigma_n^2)}$ and off-diagonal components

$-\frac{\sigma_s^2}{\sigma_n^2(ki\sigma_s^2+\sigma_n^2)}$. Similarly,

$$\sigma^2_{Z_x|d_{i+1}} = \sigma_n^2\left(1+\frac{\sigma_s^2}{k(i+1)\sigma_s^2+\sigma_n^2}\right)$$

and

$$\mu_{Z_x|d_i} = \frac{\sigma_n^2}{ki\sigma_s^2+\sigma_n^2}\mu + \frac{\sigma_s^2}{ki\sigma_s^2+\sigma_n^2}\mathbf{z}_{\mathbf{x}_{1:i}}\mathbf{1}^\top \tag{A.16}$$

where $\mu \triangleq \mu_{Z_x} = \mu_{Z_u}$ for $x, u \in \mathcal{X}$ due to the constant mean assumption (Section 3.5.1). If x is a component of $\mathbf{x}_{i+1}$ but not a component of $\mathbf{x}_{1:i}$, $\sigma^2_{Z_x|d_i}$ and $\mu_{Z_x|d_i}$ are calculated in the same way as that of (A.15) and (A.16) respectively but $\sigma^2_{Z_x|d_{i+1}} = 0$. Then,

$$\begin{aligned}
R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = &\ \textstyle\sum_{x\in\mathcal{X}} \exp\{2\mu_{Z_x|d_i} + \sigma^2_{Z_x|d_i}\}(\exp\{\sigma^2_{Z_x|d_i} - \sigma^2_{Z_x|d_{i+1}}\} - 1) \\
= &\ \exp\left\{2\left(\frac{\sigma_n^2}{ki\sigma_s^2+\sigma_n^2}\mu + \frac{\sigma_s^2}{ki\sigma_s^2+\sigma_n^2}\mathbf{z}_{\mathbf{x}_{1:i}}\mathbf{1}^\top\right) + \sigma_n^2\left(1+\frac{\sigma_s^2}{ki\sigma_s^2+\sigma_n^2}\right)\right\} \\
&\left\{|\mathcal{X}''|\left(\exp\left\{\sigma_n^2\sigma_s^2\left(\frac{1}{ki\sigma_s^2+\sigma_n^2} - \frac{1}{k(i+1)\sigma_s^2+\sigma_n^2}\right)\right\} - 1\right) + \right. \\
&\left. |\mathcal{X}'|\left(\exp\left\{\sigma_n^2\left(1+\frac{\sigma_s^2}{ki\sigma_s^2+\sigma_n^2}\right)\right\} - 1\right)\right\} \\
= &\ \exp\left\{2\left(\frac{\sigma_n^2}{ki\sigma_s^2+\sigma_n^2}\mu + \frac{\sigma_s^2}{ki\sigma_s^2+\sigma_n^2}\mathbf{z}_{\mathbf{x}_{1:i}}\mathbf{1}^\top\right) + \sigma_n^2\left(1+\frac{\sigma_s^2}{ki\sigma_s^2+\sigma_n^2}\right)\right\} \\
&\left\{|\mathcal{X}-k(i+1)|\left(\exp\left\{\sigma_n^2\sigma_s^2\left(\frac{1}{ki\sigma_s^2+\sigma_n^2} - \frac{1}{k(i+1)\sigma_s^2+\sigma_n^2}\right)\right\} - 1\right) + \right. \\
&\left. k\left(\exp\left\{\sigma_n^2\left(1+\frac{\sigma_s^2}{ki\sigma_s^2+\sigma_n^2}\right)\right\} - 1\right)\right\}.
\end{aligned} \tag{A.17}$$

where $\mathcal{X}'' = \{x \mid x \text{ is not a location component of } \mathbf{x}_{1:i+1}\}$, $|\mathcal{X}'| = k$, and $|\mathcal{X}''| = |\mathcal{X}-k(i+1)|$. Note that $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ is independent of $\mathbf{x}_{i+1}$ and is the exponential of an affine function of

$\mathbf{z}_{\mathbf{x}_{1:i}}$ only. So, we let

$$R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = \alpha_i \exp\left\{\frac{2\sigma_s^2}{ki\sigma_s^2 + \sigma_n^2}\mathbf{z}_{\mathbf{x}_{1:i}}\mathbf{1}^\top\right\}$$

where α_i is the constant term in (A.17). Consequently, we can swap the maximum and expectation in (3.14) for stage $t-1$ to give

$$\begin{aligned}
U_{t-1}^{\pi^1}(d_{t-1}) = & \max_{\mathbf{a}_{t-1}\in\mathbf{A}(\mathbf{x}_{t-1})} R^{\pi^1}(\mathbf{x}_t, d_{t-1}) + \int f(\mathbf{z}_{\mathbf{x}_t} \mid d_{t-1})\; U_t^{\pi^1}(d_t)\; \mathrm{d}\mathbf{z}_{\mathbf{x}_t} \\
= & \max_{\mathbf{a}_{t-1}} \alpha_{t-1} \exp\left\{\frac{2\sigma_s^2}{k(t-1)\sigma_s^2 + \sigma_n^2}\mathbf{z}_{\mathbf{x}_{1:t-1}}\mathbf{1}^\top\right\} + \\
& \max_{\mathbf{a}_t} \alpha_t \exp\left\{\frac{2\sigma_s^2}{kt\sigma_s^2 + \sigma_n^2}\mathbf{z}_{\mathbf{x}_{1:t-1}}\mathbf{1}^\top\right\} \mathbb{E}\left[\exp\left\{\frac{2\sigma_s^2}{kt\sigma_s^2 + \sigma_n^2}\mathbf{Z}_{\mathbf{x}_t}\mathbf{1}^\top\right\} \middle| d_{t-1}\right] \\
= & \max_{\mathbf{a}_{t-1:t}} \alpha_{t-1} \exp\left\{\frac{2\sigma_s^2}{k(t-1)\sigma_s^2 + \sigma_n^2}\mathbf{z}_{\mathbf{x}_{1:t-1}}\mathbf{1}^\top\right\} + \\
& \alpha_t \exp\left\{\frac{2\sigma_s^2}{kt\sigma_s^2 + \sigma_n^2}\left(\mathbf{z}_{\mathbf{x}_{1:t-1}}\mathbf{1}^\top + \boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_t}|d_{t-1}}\mathbf{1}^\top + \frac{\sigma_s^2}{kt\sigma_s^2 + \sigma_n^2}\mathbf{1}\Sigma_{\mathbf{Z}_{\mathbf{x}_t}|d_{t-1}}\mathbf{1}^\top\right)\right\}
\end{aligned} \tag{A.18}$$

where $\boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_t}|d_{t-1}}$ and $\Sigma_{\mathbf{Z}_{\mathbf{x}_t}|d_{t-1}}$ are independent of $\mathbf{x}_{t:t+1}$, and $\boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_t}|d_{t-1}}$ is an affine function of $\mathbf{z}_{\mathbf{x}_{1:t-1}}$ only. So, the expression after the maximum operator in the third equality of (A.18) is independent of $\mathbf{x}_{t:t+1}$ and is the sum of exponentials of affine function of $\mathbf{z}_{\mathbf{x}_{1:t-1}}$ only. As a result, we can swap the maximum and expectation in (3.14) for stage $t-2$ to derive an expression after the maximum operator that is independent of $\mathbf{x}_{t-1:t+1}$ and is the sum of exponentials of affine function of $\mathbf{z}_{\mathbf{x}_{1:t-2}}$ only. We continue swapping the maximum and expectation in (3.14) for stages $t-3, \ldots, 0$ until the single-staged equation of MASP(n) (3.16) is obtained. So, $\pi^1 = \pi^n$, which is independent of $\mathbf{z}_{\mathbf{x}_{1:t+1}}$. Hence, π^1 is non-adaptive.

A.14.2 iMASP

Recall that the reward function is

$$R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = \mathbb{H}[\mathbf{Y}_{\mathbf{x}_{i+1}} \mid d_i] = \log \sqrt{(2\pi e)^k |\Sigma_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i}|} + \boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mathbf{1}^\top \ .$$

Case of $\ell = 0$: $R^{\pi^1}(\mathbf{x}_{i+1}, d_i) = k \left\{ \frac{1}{2} \log 2\pi e(\sigma_s^2 + \sigma_n^2) + \mu \right\}$. Hence, Lemma 3.5.1 is satisfied. As a result, Theorem 3.5.1 holds. Therefore, π^1 is non-adaptive. The reduction of iMASP(1) to a single-staged iMASP(n) is similar to that of (3.22).

Case of $\ell = \infty$: To simplify exposition, we assume that no prior data d_0 are available. Then,

$$\boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{i+1}}|d_i} \mathbf{1}^\top = \sum_{x \in \mathcal{X}'} \mu_{Z_x|d_i} = k \left(\frac{\sigma_n^2}{ki\sigma_s^2 + \sigma_n^2} \mu + \frac{\sigma_s^2}{ki\sigma_s^2 + \sigma_n^2} \mathbf{z}_{\mathbf{x}_{1:i}} \mathbf{1}^\top \right)$$

where $\mu_{Z_x|d_i}$ is calculated using (A.16). So, $R^{\pi^1}(\mathbf{x}_{i+1}, d_i)$ is independent of $\mathbf{x}_{i+1}$ and is an affine function of $\mathbf{z}_{\mathbf{x}_{1:i}}$ only. Consequently, we can swap the maximum and expectation in (3.14) for stage $t-1$. The rest of the proof follows closely to that for MASP in Section A.14.1, and the single-staged equation of MASP(n) (3.16) is consequently derived. So, $\pi^1 = \pi^n$, which is independent of $\mathbf{z}_{\mathbf{x}_{1:t+1}}$. Hence, π^1 is non-adaptive.

A.15 Equation 8.2

Using Lemma 8.2.1,

$$\begin{aligned} \mathbb{H}[Y_{x_2} \mid d_1] = & \ \frac{1}{2} \log 2\pi e \sigma^2_{Z_{x_2}|d_1} + \mu_{Z_{x_2}|d_1} \\ = & \ \frac{1}{2} \log 2\pi e \sigma^2_{Z_{x_2}|d_1} + (1 - w_0(x_2) - w_1(x_2)) \mu_{Z_{x_2}} + w_0(x_2) z_{x_0} + w_1(x_2) z_{x_1} \ , \end{aligned}$$

which is a linear function of z_{x_1}. Then,

$$\begin{aligned}
&\int_{\mathcal{Z}_{x_1}^{[i]}} f(z_{x_1} \mid d_0)\ \mathbb{H}[Y_{x_2^{[i]}} \mid d_1]\ dz_{x_1} \\
&= \int_{\mathcal{Z}_{x_1}^{[i]}} \frac{1}{\sqrt{2\pi}\sigma_{Z_{x_1}|d_0}} \exp\left\{-\frac{(z_{x_1}-\mu_{Z_{x_1}|d_0})^2}{2\sigma^2_{Z_{x_1}|d_0}}\right\}\left(c(d_0,x_1,x_2^{[i]}) + w_1(x_2^{[i]})z_{x_1}\right)\ dz_{x_1} \\
&= \frac{c(d_0,x_1,x_2^{[i]}) + w_1(x_2^{[i]})\mu_{Z_{x_1}|d_0}}{2}\left[\operatorname{erf}\left(\frac{z_{x_1}-\mu_{Z_{x_1}|d_0}}{\sqrt{2}\sigma_{Z_{x_1}|d_0}}\right)\right]_{\mathcal{Z}_{x_1}^{[i]}} - \\
&\quad \frac{w_1(x_2^{[i]})\sigma_{Z_{x_1}|d_0}}{\sqrt{2\pi}}\left[\exp\left\{-\frac{(z_{x_1}-\mu_{Z_{x_1}|d_0})^2}{2\sigma^2_{Z_{x_1}|d_0}}\right\}\right]_{\mathcal{Z}_{x_1}^{[i]}}
\end{aligned} \tag{A.19}$$

where $c(d_0,x_1,x_2) = \frac{1}{2}\log 2\pi e\sigma^2_{Z_{x_2}|d_1} + (1-w_0(x_2)-w_1(x_2))\mu_{Z_{x_2}} + w_0(x_2)z_{x_0}$.

A.16 Lemma 9.3.1

Let $\Sigma_{\mathbf{x}_{0:i-1}\mathbf{x}_{0:i-1}|x_i} \stackrel{\triangle}{=} \mathbf{C}+\mathbf{E}$ where $\mathbf{C}$ is defined to be a matrix with diagonal components $\sigma^2_{Z_{x_k}} = \sigma_s^2+\sigma_n^2$ for $k=0,\ldots,i-1$ and off-diagonal components 0, and $\mathbf{E}$ is a matrix with diagonal components $-(\sigma_{Z_{x_k}Z_{x_i}})^2/\sigma^2_{Z_{x_i}} = -(\sigma_{Z_{x_k}Z_{x_i}})^2/(\sigma_s^2+\sigma_n^2)$ for $k=0,\ldots,i-1$ and the same off-diagonal components as $\Sigma_{\mathbf{x}_{0:i-1}\mathbf{x}_{0:i-1}|x_i}$ (i.e., $\sigma_{Z_{x_j}Z_{x_k}|x_i} = \sigma_{Z_{x_j}Z_{x_k}} - \sigma_{Z_{x_j}Z_{x_i}}\sigma_{Z_{x_i}Z_{x_k}}/\sigma^2_{Z_{x_i}}$ for $j,k=0,\ldots,i-1$, $j\neq k$). Then,

$$||\mathbf{C}^{-1}||_2 = ||(\sigma_s^2+\sigma_n^2)^{-1}\mathbf{I}||_2 = \frac{1}{\sigma_s^2+\sigma_n^2}\ . \tag{A.20}$$

The last equality follows from $\sigma_s^2+\sigma_n^2$ being the smallest eigenvalue of $\mathbf{C}$. So, $1/(\sigma_s^2+\sigma_n^2)$ is the largest eigenvalue of $\mathbf{C}^{-1}$, which is equal to $||\mathbf{C}^{-1}||_2$.

Note that the minimum distance between any pair of location components of $\mathbf{x}_{0:i-1}$ cannot be less than ω_x. So, it can be observed that any component of $\mathbf{E}$ cannot have an absolute

value more than $\sigma_s^2\xi$. Therefore,

$$||\mathbf{E}||_2 \leq i\sigma_s^2\xi \ , \tag{A.21}$$

which follows from a property of the matrix 2-norm that $||\mathbf{E}||_2$ cannot be more than the largest absolute component of $\mathbf{E}$ multiplied by i (Golub and Van Loan, 1996).

Note that the minimum distance between locations x_i and x_{i+1} as well as between location x_i and any location component of $\mathbf{x}_{0:i-1}$ cannot be less than ω_x. So, it can be observed that any component of $\Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}$ cannot have an absolute value more than $\sigma_s^2\xi^2$. Therefore,

$$|\sigma_{Z_{x_{i+1}}Z_{x_k}|x_i}| \leq \sigma_s^2\xi^2 \tag{A.22}$$

for $k = 0, \ldots, i-1$.

Now,

$$\begin{aligned}
&\Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\Sigma^{-1}_{\mathbf{x}_{0:i-1}\mathbf{x}_{0:i-1}|x_i}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} - \Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\Sigma^{-1}_{\mathbf{x}_{0:i-1}\mathbf{x}_{0:i-1}|x_i}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} \\
&= \Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}(\mathbf{C}+\mathbf{E})^{-1}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} - \Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\mathbf{C}^{-1}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} \\
&= \Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\{(\mathbf{C}+\mathbf{E})^{-1} - \mathbf{C}^{-1}\}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} \\
&\leq ||\Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}||_2^2 \ ||(\mathbf{C}+\mathbf{E})^{-1} - \mathbf{C}^{-1}||_2 \\
&\leq \sum_{k=0}^{i-1} |\sigma_{Z_{x_{i+1}}Z_{x_k}|x_i}|^2 \frac{||\mathbf{C}^{-1}||_2 \ ||\mathbf{E}||_2}{\frac{1}{||\mathbf{C}^{-1}||_2} - ||\mathbf{E}||_2} \\
&= i(\sigma_s^2)^2\xi^4 \frac{||\mathbf{C}^{-1}||_2 \ ||\mathbf{E}||_2}{\frac{1}{||\mathbf{C}^{-1}||_2} - ||\mathbf{E}||_2} \ .
\end{aligned} \tag{A.23}$$

The first inequality is due to Cauchy-Schwarz inequality and submultiplicativity of the matrix norm (Stewart and Sun, 1990). The second inequality follows from an important result in the perturbation theory of matrix inverses (in particular, Theorem III.2.5 in (Stewart and Sun, 1990)). It requires the assumption of $||\mathbf{C}^{-1} \ \mathbf{E}||_2 < 1$. This assumption can be satisfied by $||\mathbf{C}^{-1}||_2 \ ||\mathbf{E}||_2 < 1$ because $||\mathbf{C}^{-1} \ \mathbf{E}||_2 \leq ||\mathbf{C}^{-1}||_2 \ ||\mathbf{E}||_2$. By (A.20) and (A.21), $||\mathbf{C}^{-1}||_2 \ ||\mathbf{E}||_2 < 1$ translates to $\xi < \rho/i$. The last equality is due to (A.22).

From (A.23),

$$
\begin{aligned}
&\Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}(\mathbf{C}+\mathbf{E})^{-1}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} \\
&\leq \Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\mathbf{C}^{-1}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} + i(\sigma_s^2)^2\xi^4\frac{||\mathbf{C}^{-1}||_2\ ||\mathbf{E}||_2}{\frac{1}{||\mathbf{C}^{-1}||_2}-||\mathbf{E}||_2} \\
&\leq i(\sigma_s^2)^2\xi^4\ ||\mathbf{C}^{-1}||_2\left(1+\frac{||\mathbf{E}||_2}{\frac{1}{||\mathbf{C}^{-1}||_2}-||\mathbf{E}||_2}\right) \\
&= \frac{i(\sigma_s^2)^2\xi^4}{\frac{1}{||\mathbf{C}^{-1}||_2}-||\mathbf{E}||_2} \\
&\leq \frac{i(\sigma_s^2)^2\xi^4}{\frac{1}{||\mathbf{C}^{-1}||_2}-||\mathbf{E}||_2} \\
&\leq \frac{i(\sigma_s^2)^2\xi^4}{\sigma_s^2+\sigma_n^2-i\sigma_s^2\xi} \\
&= \frac{\sigma_s^2\xi^4}{\frac{\rho}{i}-\xi}
\end{aligned}
\tag{A.24}
$$

The second inequality is due to

$$
\Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\mathbf{C}^{-1}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} \leq i(\sigma_s^2)^2\xi^4\ ||\mathbf{C}^{-1}||_2\ ,
$$

which follows from Cauchy-Schwarz inequality and (A.22). The third inequality follows from (A.20) and (A.21).

We will need the following property of posterior variance, which is similar to (3.20):

$$
\sigma^2_{Z_{x_{i+1}}|\mathbf{x}_{0:i}} = \sigma^2_{Z_{x_{i+1}}|x_i} - \Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\Sigma^{-1}_{\mathbf{x}_{0:i-1}\mathbf{x}_{0:i-1}|x_i}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} \tag{A.25}
$$

where $\Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}$ is a posterior covariance vector with components $\sigma_{Z_{x_{i+1}}Z_{x_k}|x_i}$ for $k = 0,\ldots,i-1$, $\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i}$ is the transpose of $\Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}$, and $\Sigma_{\mathbf{x}_{0:i-1}\mathbf{x}_{0:i-1}|x_i}$ is a posterior covariance matrix with components $\sigma_{Z_{x_j}Z_{x_k}|x_i}$ for $j,k = 0,\ldots,i-1$.

By (A.24) and (A.25),

$$\begin{aligned}\sigma^2_{Z_{x_{i+1}}|x_i} - \sigma^2_{Z_{x_{i+1}}|\mathbf{x}_{0:i}} &= \Sigma_{x_{i+1}\mathbf{x}_{0:i-1}|x_i}\Sigma^{-1}_{\mathbf{x}_{0:i-1}\mathbf{x}_{0:i-1}|x_i}\Sigma_{\mathbf{x}_{0:i-1}x_{i+1}|x_i} \\ &\leq \frac{\sigma_s^2\xi^4}{\frac{\rho}{i}-\xi} \ .\end{aligned}$$

A.17 Theorem 9.3.2

Proof by induction on i that $U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq \widetilde{U}_i(x_i) \leq U_i^{\pi^1}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t}\Delta(s)$ for $i = t, \ldots, 0$.

Base case ($i = t$): By Lemma 9.3.2,

$$\begin{aligned}&\mathbb{H}[Z_{x_{t+1}} \mid \mathbf{x}_{0:t}] \leq \mathbb{H}[Z_{x_{t+1}} \mid x_t] \leq \mathbb{H}[Z_{x_{t+1}} \mid \mathbf{x}_{0:t}] + \Delta(t) \quad \text{for any } x_{t+1} \\ \Rightarrow\ &\max_{a_t\in\mathcal{A}(x_t)} \mathbb{H}[Z_{x_{t+1}} \mid \mathbf{x}_{0:t}] \leq \max_{a_t\in\mathcal{A}(x_t)} \mathbb{H}[Z_{x_{t+1}} \mid x_t] \leq \max_{a_t\in\mathcal{A}(x_t)} \mathbb{H}[Z_{x_{t+1}} \mid \mathbf{x}_{0:t}] + \Delta(t) \qquad \text{(A.26)} \\ \Rightarrow\ &U_t^{\pi^1}(\mathbf{x}_{0:t}) \leq \widetilde{U}_t(x_t) \leq U_t^{\pi^1}(\mathbf{x}_{0:t}) + \Delta(t) \ .\end{aligned}$$

Hence, the base case is true.

Inductive case: Suppose that

$$U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) \leq \widetilde{U}_{i+1}(x_{i+1}) \leq U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) + \sum_{s=i+1}^{t}\Delta(s) \qquad \text{(A.27)}$$

is true. We have to prove that $U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq \widetilde{U}_i(x_i) \leq U_i^{\pi^1}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t}\Delta(s)$ is true.

We will first show that $\widetilde{U}_i(x_i) \leq U_i^{\pi^1}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s)$. By Lemma 9.3.2,

$$
\begin{aligned}
& \mathbb{H}[Z_{x_{i+1}} \mid x_i] \leq \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] + \Delta(i) \quad \text{for any } x_{i+1} \\
\Rightarrow \quad & \mathbb{H}[Z_{x_{i+1}} \mid x_i] + \widetilde{U}_{i+1}(x_{i+1}) \leq \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) + \sum_{s=i}^{t} \Delta(s) \text{ by (A.27) for any } x_{i+1} \\
\Rightarrow \quad & \max_{a_i \in \mathcal{A}(x_i)} \mathbb{H}[Z_{x_{i+1}} \mid x_i] + \widetilde{U}_{i+1}(x_{i+1}) \leq \max_{a_i \in \mathcal{A}(x_i)} \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) + \sum_{s=i}^{t} \Delta(s) \\
\Rightarrow \quad & \widetilde{U}_i(x_i) \leq U_i^{\pi^1}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s) \ .
\end{aligned}
$$

We will now prove that $U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq \widetilde{U}_i(x_i)$. By Lemma 9.3.2,

$$
\begin{aligned}
& \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] \leq \mathbb{H}[Z_{x_{i+1}} \mid x_i] \quad \text{for any } x_{i+1} \\
\Rightarrow \quad & \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) \leq \mathbb{H}[Z_{x_{i+1}} \mid x_i] + \widetilde{U}_{i+1}(x_{i+1}) \text{ by (A.27) for any } x_{i+1} \\
\Rightarrow \quad & \max_{a_i \in \mathcal{A}(x_i)} \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) \leq \max_{a_i \in \mathcal{A}(x_i)} \mathbb{H}[Z_{x_{i+1}} \mid x_i] + \widetilde{U}_{i+1}(x_{i+1}) \\
\Rightarrow \quad & U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq \widetilde{U}_i(x_i) \ .
\end{aligned}
$$

Hence, the inductive case is true.

A.18 Theorem 9.3.3

The following lemma is needed for the proof:

Lemma A.18.1. $\widetilde{U}_i(x_i) \leq U_i^{\widetilde{\pi}}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s)$ for $i = 0, \ldots, t$.

Proof by induction on i that $U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq U_i^{\widetilde{\pi}}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s)$ for $i = t, \ldots, 0$.

Base case ($i = t$):

$$U_t^{\pi^1}(\mathbf{x}_{0:t}) \leq \widetilde{U}_t(x_t) \leq U_t^{\widetilde{\pi}}(\mathbf{x}_{0:t}) + \Delta(t) \ .$$

The first inequality is due to Theorem 9.3.2. The second inequality follows from Lemma A.18.1. Hence, the base case is true.

Inductive case: Suppose that

$$U_{i+1}^{\pi^1}(\mathbf{x}_{0:i+1}) \leq U_{i+1}^{\widetilde{\pi}}(\mathbf{x}_{0:i+1}) + \sum_{s=i+1}^{t} \Delta(s) \tag{A.28}$$

is true. We have to prove that $U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq U_i^{\widetilde{\pi}}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s)$ is true.

$$\begin{aligned}
U_i^{\pi^1}(\mathbf{x}_{0:i}) \leq \ & \widetilde{U}_i(x_i) \\
= \ & \mathbb{H}[Z_{\tau(x_i,\widetilde{\pi}_i(x_i))} \mid x_i] + \widetilde{U}_{i+1}(\tau(x_i, \widetilde{\pi}_t(x_i))) \\
\leq \ & \mathbb{H}[Z_{\tau(x_i,\widetilde{\pi}_i(\mathbf{x}_{0:i}))} \mid \mathbf{x}_{0:i}] + \Delta(i) + \widetilde{U}_{i+1}(\tau(x_i, \widetilde{\pi}_t(x_i))) \\
\leq \ & \mathbb{H}[Z_{\tau(x_i,\widetilde{\pi}_i(\mathbf{x}_{0:i}))} \mid \mathbf{x}_{0:i}] + \Delta(i) + U_{i+1}^{\widetilde{\pi}}(\mathbf{x}_{0:i+1}) + \sum_{s=i+1}^{t} \Delta(s) \\
= \ & U_i^{\widetilde{\pi}}(\mathbf{x}_{0:i}) + \textstyle\sum_{s=i}^{t} \Delta(s) \ .
\end{aligned}$$

The first inequality is due to Theorem 9.3.2. The first equality follows from (9.4). The second inequality follows from Lemma 9.3.2. The third inequality is due to Lemma A.18.1. Hence, the inductive case is true.

A.19 Lemma A.18.1

Proof by induction on i that $\widetilde{U}_i(x_i) \leq U_i^{\widetilde{\pi}}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s)$ for $i = t, \ldots, 0$.

Base case ($i = t$):

$$\begin{aligned}
\widetilde{U}_t(x_t) = & \max_{a_t \in \mathcal{A}(x_t)} \mathbb{H}[Z_{x_{t+1}} \mid x_t] \\
= & \ \mathbb{H}[Z_{\tau(x_t, \widetilde{\pi}_t(x_t))} \mid x_t] \\
\leq & \ \mathbb{H}[Z_{\tau(x_t, \widetilde{\pi}_t(\mathbf{x}_{0:t}))} \mid \mathbf{x}_{0:t}] + \Delta(t) \\
= & \ U_t^{\widetilde{\pi}}(\mathbf{x}_{0:t}) + \Delta(t) \ .
\end{aligned}$$

The first equality follows from (9.4). The inequality follows from Lemma 9.3.2. The last equality is due to (9.7). So, the base case is true.

Inductive case: Suppose that

$$\widetilde{U}_{i+1}(x_{i+1}) \leq U_{i+1}^{\widetilde{\pi}}(\mathbf{x}_{0:i+1}) + \sum_{s=i+1}^{t} \Delta(s) \tag{A.29}$$

is true. We have to prove that $\widetilde{U}_i(x_i) \leq U_i^{\widetilde{\pi}}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s)$ is true.

By Lemma 9.3.2,

$$\begin{aligned}
& \mathbb{H}[Z_{x_{i+1}} \mid x_i] \leq \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] + \Delta(i) \quad \text{for any } x_{i+1} \\
\Rightarrow \ & \mathbb{H}[Z_{x_{i+1}} \mid x_i] + \widetilde{U}_{i+1}(x_{i+1}) \leq \mathbb{H}[Z_{x_{i+1}} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\widetilde{\pi}}(\mathbf{x}_{0:i+1}) + \sum_{s=i}^{t} \Delta(s) \text{ by (A.29) for any } x_{i+1} \\
\Rightarrow \ & \mathbb{H}[Z_{\tau(x_i, \widetilde{\pi}_i(x_i))} \mid x_i] + \widetilde{U}_{i+1}(\tau(x_i, \widetilde{\pi}_t(x_i))) \leq \mathbb{H}[Z_{\tau(x_i, \widetilde{\pi}_i(\mathbf{x}_{0:i}))} \mid \mathbf{x}_{0:i}] + U_{i+1}^{\widetilde{\pi}}(\mathbf{x}_{0:i+1}) + \sum_{s=i}^{t} \Delta(s) \\
& \text{for } x_{i+1} = \tau(x_i, \widetilde{\pi}_t(x_i)) = \tau(x_i, \widetilde{\pi}_i(\mathbf{x}_{0:i})) \\
\Rightarrow \ & \widetilde{U}_i(x_i) \leq U_i^{\widetilde{\pi}}(\mathbf{x}_{0:i}) + \sum_{s=i}^{t} \Delta(s) \ .
\end{aligned}$$

Hence, the inductive case is true.