Efficiently Sampling from Underlying Physical Models

Greydon Taylor Foil

CMU-RI-TR-16-57

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Robotics

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

October 12, 2016

Thesis Committee:
David Wettergreen, Chair
George Kantor
Jeff Schneider
David R. Thompson, Jet Propulsion Laboratory

Abstract

The capability and mobility of exploration robots is increasing rapidly, yet missions will always be constrained by one main resource: time. Time limits the number of samples a robot can collect, sites it can analyze, and the availability of human oversight, so it is imperative the robot is able to make intelligent actions when it comes to choosing when, where, and what to sample, a process known as *adaptive sampling*.

Current Mars rover operations give an example of the need for adaptive sampling techniques. Daily plans budget specific operations down to the minute, and each operation is given a maximum of a few hours to complete. However, the capabilities of the rover far exceed the limited plans. For example, the Mars 2020 mission is testing the Planetary Instrument for X-Ray Lithochemistry (PIXL), a spectrometer used to analyze rocks on a microscopic scale. A quick scan of a postage stamp-sized surface can take an hour but is very noisy, yet reducing this noise with a full scan of the surface takes multiple days. Scheduling constraints and communication delay prohibits scientists from reviewing data in real-time and they are thus unable to follow up at any promising sample locations. However, adaptive sampling techniques enable the rover to quickly and autonomously choose a subset of interesting points for detailed follow-up scans.

This work advances the state of the art in adaptive sampling for exploration robotics. We take advantage of the fact that rover operations are typically not performed in a vacuum; extensive contextual data is often present, most often in the form of orbital imagery, rover camera images, and quick microscopic scans like those described above. Using this context, we apply advanced Bayesian and nonparametric models to decide where best to sample under a limited budget. Unlike previous works, our approaches reduce the impact of noise on sample site selection, a common problem when using contextual data.

The thesis evaluates our methods in three main scenarios. We begin with the general case, in which noisy contextual information of an entire scene is available to the rover. The rover must choose sampling locations that are expected to create a maximally diverse sample set or a sample set that best represents the entire scene. Sampling a point reveals the true underlying data, altering the reward of future points.

Next, we demonstrate improvements to adaptive sampling techniques performed at the microscopic scale, as if the rover stopped for a detailed inspection. Mimicking PIXL operations, the rover no longer has full contextual information, but instead collects a quick scan of each point, one at a time, and decides whether or not to perform a full scan at that point before moving on. We demonstrate a Dirichlet-based technique for building a classification of samples, and show improved performance over existing sampling methods.

Finally, we consider rover operations the orbital level, demonstrating improvements to rover path selection using satellite maps as contextual data. Here we apply our Bayesian models to maximize the expected yield of samples collected across a number of prospective sampling paths, ultimately choosing the path most likely to maximize our understanding of the terrain.

II

Acknowledgments

This thesis was a huge amount of work and much of it was only possible through the help of others. I would like to thank the other students and staff of the Robotics Institute and the Field Robotics Center, for their help and cooperation on a number of projects. Specifically, I would like to thank the Life in the Atacama staff and affiliates: Srinivasan Vijayarangan, Jim Teza, Mike Wagner, Michael Furlong, and Dave Kohanbash. Through our work together I have learned much about the practical realities of field robotics, tricks learned over years of experience, and how ridiculous field testing can sometimes be (eg. flash floods in the driest desert on earth?!).

Our orbital experiments relied on the AVIRIS-NG and simulated CRISM that would not be possible without the help of Alberto Candela, Eldar Noe Dobrea, and David Thompson.

I would like to thank the staff of the Intelligent Robotics Group at NASA Ames Research Center, and the NSTRF program for invaluable experiences over the past few years. My NSTRF scholarship allowed me to travel and work with multiple groups at the forefront of planetary robotics, including the staff at the IRG, and I am very grateful for the experience. I would specifically like to thank Terry Fong, my mentor in the NSTRF program, for his research suggestions and the large amount of networking he went out of his way to coordinate during my visits to Ames.

Thanks to George Kantor and Jeff Schneider, my committee members, for offering consultations and advice throughout the thesis process. It is easy to get wrapped up in the world of planetary robotics, and their alternative views on agricultural robotics and statistical theory were important in keeping my focus broad.

I would also like to extend a huge thanks to my external committee member, David Thompson. He has twice given me opportunities to work at the Jet Propulsion Laboratory with some of the people on the forefront of planetary robotics, for which I am very grateful. His experience, knowledge, and capabilities in planetary robotics, machine learning, and remote sensing are unmatched, and I thank him for answering the thousands of questions I have asked over the years.

The person I owe greatest thanks to is surely my advisor, David Wettergreen. I would like to thank him for his support throughout the years, his never-failing positive attitude, and gentle nudges in the right direction. His intuition has led me to solve a number of challenging and important problems in my research, and I am greatly thankful for his advice over the years.

Finally, I would like to thank my family, especially my wife Lynn, for their support and encouragement through the previous years and giving me the flexibility to uproot our lives, move to Pittsburgh, and return to school after years of industry.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As remote exploration robots increase in capability and traversability, missions are measured less by "what" a robot can do and more by "how", "why", and "where" a robot can act to make the most of its mission objectives. We have moved away from asking questions like "Can we build a robot to climb into a volcano" or "Can we land a rover on Venus" and now ask questions like "How do we make the most of a robot driving on Mars for a year or more?" Time is the constraining resource. In order to maximize the return on a mission that time must be used wisely.

Since many robotic missions do not have constant human oversight, technologies to improve a robot's understanding and interpretation of its environment are increasingly important. Instead of waiting for human intervention, robots that understand their environment can act on their own in intelligent ways, making the most of their time.

Adaptive sampling is one way in which a rover might improve its utilization of its time. Adaptive sampling refers to techniques that use onboard models and statistical analysis to choose where and how to deploy instruments and acquire measurements. Adaptive sampling methods might be used to detect anomalous features, plan paths with stops at diverse types of terrain, or analyze an interesting sample and decide to take more detailed measurements. These techniques can increase data acquisition rates and quality, improving the overall yield of a mission.

In order to make the most of our robot's capabilities, it is necessary to develop adaptive sampling techniques that optimize the quality of the sampling data collected by a robot while simultaneously allowing for greater exploration. Robots have become very good at detecting objects directly in front of them [Foil et al., 2013, Pedersen et al., 2001, Thompson, 2008, Hwangbo et al., 2009], environmental phenomena that passes within sight of their cameras [Estlin et al., 2012], or finding approximate measurement locations only using low-resolution orbital imagery [Thompson et al., 2015b]. However, these methods are incomplete or only run under certain conditions, causing the rover to miss interesting objects or measurement locations during a traverse.

Figure 1.1 illustrates recent difficulties encountered during field tests in the Atacama Desert in Chile and motivations for this work. In these images we see examples of problems caused by the myopic world view presented by local navigation cameras. If orbital sensing does not

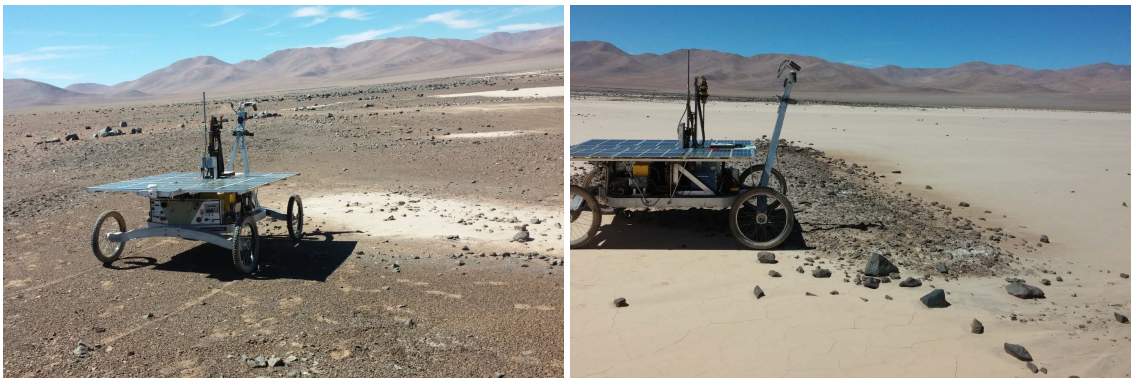


Figure 1.1: Two instances of where better target selection could have improved Zoë's science return. In both situations Zoë navigates to a location and collects a grid of spectra immediately in front of it. In the left image Zoë has correctly identified a region of interest and navigated there, but is pointing in an incorrect direction and samples from a mixture of the materials. In the right image Zoë has again navigated to the correct location but aims its spectrometer beyond the optimal targets, including patches of white salts that are of interest to the science team.

reveal a target of interest, or if the robot is unable to navigate to a target location with sufficient accuracy, the rover is only able to observe what is directly in front of it, leading to samples with poor information content. Without better onboard models and sampling strategies, many informative sampling opportunities are overlooked during the traverse.

In the cases shown here in the Atacama Desert, the rover has driven to the correct location, but upon arrival is pointing in the wrong direction. As it approached the target location, both sampling targets were visible in camera imagery, but that information was not acted upon and the rover missed sampling the targets of interest entirely. Situations such as these present a myriad of opportunities for improvement, through both terrain understanding and adaptive sampling techniques.

The overall objective of this work is to introduce adaptive sampling methods that increase the diversity of samples collected and collect representative samples. We accomplish this goal by making use of rover measurements that can be gathered quickly or with little effort, such as images or rapid contextual spectral scans, to create estimates of models of the domain, then sampling from these models in order to confirm measurements we expect to see and spot interesting areas that deviate from our models.

This work maximizes two objectives in an *efficient* manner: sample *diversity* and collecting *representative* samples. Here efficiency is defined as the amount of information gained per unit of time. This is maximized by focusing observation on areas expected to improved our knowledge and not oversampling areas we already understand. We define diversity as a measure of the different types of samples collected, aiming to collect the most widespread set of samples possible in a minimal number of measurements. Representative samples are those that are characteristic of the sample space. These samples provide the most information about the composition sample space.

These objectives are motivated by the goals of the upcoming Mars 2020 mission. Two of the

main objectives of this mission are searching for evidence of past life on Mars and collecting samples for possible return to Earth on a future mission. Signs of life will be detected by evaluating regions for habitability and preservation, then sampling within these regions, targeting materials with high biosignature preservation potential. Biosignatures themselves are likely to be microscopic in size, and sampling methods must be able to adaptively search for signatures similar to those seen on Earth, as well as interesting and unique features that may be unexpected. This requires both representative sampling approaches to characterize a sampling site, as well as diverse sampling approaches that recognize unique signatures.

Sample return further motivates the development of both diverse and representative sampling techniques. The aim of this objective is to collect samples "that represent the geologic diversity of the field site" [NASA, 2016d]. These samples are likely to be chosen from a set of diverse measurement locations that characterize the region. This motivates sampling methods that choose a set of measurement locations that are expected to offer diversity, as well as methods that select samples that are representative of these locations.

Our work will demonstrate two main approaches for adaptive sampling that improve sample diversity in an efficient manner. Formally:

In robotic exploration, interpreting contextual measurements yields better science results by both increasing sampling efficiency and improving a robots ability to collect diverse and representative samples.

1.1 Adaptive Sampling: Making Informed Decisions Based on Prior Actions

At the most basic level, adaptive sampling is applicable to any scenario in which collecting a sample provides actionable information. If a robot is not acting out a prescribed sampling plan, and instead is observing, updating an internal model or estimate of sample reward, then acting on this update, adaptive sampling methods can be used. These adaptive methods can improve the sample selection process by selecting samples that improve the understanding of the sampling location.

In years past, especially with regards to planetary robotics, information collected by rovers was analyzed solely by humans, such as science teams observing robot measurements and selecting sampling targets or operators navigating robots to unknown or informative regions. As robots increase in capability and autonomy, adaptive sampling methods become more important, allowing robots to prioritize measurements for scientist review, choose high value targets when human oversight is unavailable, and increase the number of samples collected within a fixed mission budget.

Here we highlight some of the fields of robotics that motivate and can benefit from this work. These domains are not limited to just planetary robotics, but span all fields of robotics. In fact, due to rapid advancements in robotics technology, leading to numerous low-cost and capable mobile robotics platforms, adaptive sampling methods are likely to be most applicable to terrestrial robotics in the near future.

1.1.1 Planetary and Exploration Robotics

Planetary robotics missions have a number of opportunities to use adaptive sampling methods. The goals of the upcoming Mars 2020 mission include both characterizing sample sites to determine habitability, as well as the search for samples that represent the geologic diversity of sites [NASA, 2016d]. These motivate the development of adaptive sampling techniques that can characterize regions, or find representative samples, as well as techniques for maximizing the diversity of samples.

With regards to contextual data, planetary missions often make use of orbital maps for operations planning [Foil et al., 2016], localization [Hwangbo et al., 2009], and sample site selection [Thompson et al., 2015b]. It is unlikely that any upcoming planetary missions will occur without orbital imaging support. Orbital data provides powerful contextual information at the cost of resolution; most visual orbital data is approximately one meter per pixel, while multispectral data is often 15 meters per pixel or more.

This data is often insufficient for sampling purposes, but can be used as a prior for detection and prioritization of measurement locations. By conditioning on this prior data, we demonstrate adaptive sampling methods that are able to generate sparse sets of measurement locations that minimize the time a rover spends sampling, increases diversity in the data returned, while still allowing for longer traverses.

1.1.2 Microscopic Sampling

Recent planetary rovers are limited in their ability to autonomously drive and explore on the meter scale, but they are equipped with a myriad of sensing instruments that operate on the microscopic scale that allow them to collect information about their surroundings. Typically, in today's operations paradigm, the targets and placement of these instruments are hand-picked through an extensive vetting process by scientists on Earth. Candidate measurement locations are selected using visual imagery collected by the rover and returned to Earth, then follow-up operations at candidate sites are determined by scientists. If a site is likely to provide interesting scientific content, select instruments from the rover are used to prepare the surface for sampling or collect detailed measurements.

This cycle of a rover collecting measurements, transmitting them to Earth, data being reviewed by a panel of scientists, and operations planning is slow and scales poorly as more robots come into use. Communications turnaround between the rover and ground teams can take hours or a full day, leading to time wasted by the rover as it waits for instructions.

Additionally, newer instruments, such as the Planetary Instrument for X-Ray Lithochemistry (PIXL), are less amenable to this operational cycle. PIXL is an X-ray fluorescence spectrometer that collects a raster of a microscopic surface, an example of which is shown in Figure 1.3. It can collect a quick, noisy spectrum in seconds, or integrate measurements for minutes to collect a longer, more accurate spectrum. However, due to mechanical constraints, PIXL is unable to return to a specific point with a high enough accuracy to guarantee the location is in the instrument's

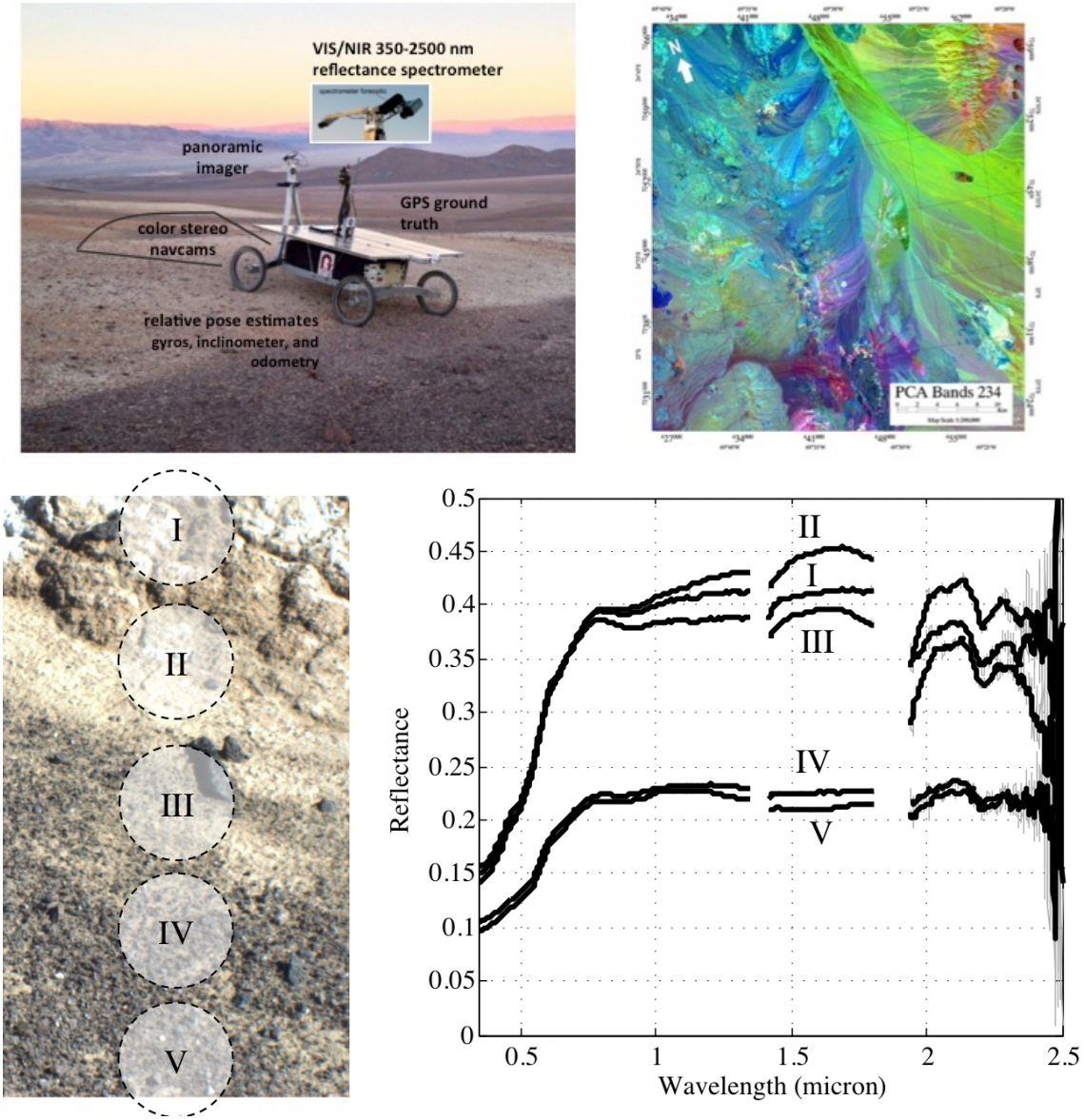
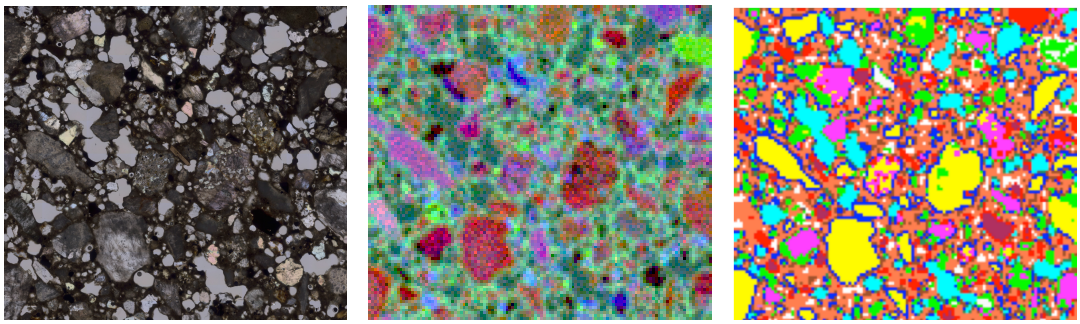


Figure 1.2: Top left: Zoë, the rover used in many of our experiments, is equipped with color navcams and a directed color camera and spectrometer. Color navcam images are inexpensive to collect, but targeted spectra each take between 5 and 150 seconds to collect, depending on their integration time. Top right: an orbital map displaying the local spectral diversity. The map displays the second, third, and fourth principal components of a multispectral ASTER image of the Atacama. Bottom: an example of spectra collected by Zoë during a traverse.



(a) Image of the surface (b) Elemental abundances shown in false color (c) Classifications

Figure 1.3: Example data from the Planetary Instrument for X-Ray Lithochemistry (PIXL) instrument. Left: A microscopic image of the surface of the "Troughite" sample. Center: A false-color image showing the responses of the elements Al (red), Ca (green), and Ti (blue). Right: A classification of the surface by clustering the spectra at each point. From [Thompson et al., 2015a].

best field of view, so it must collect a quick spectrum and decide whether or not the content of that measurement warrants follow-up. If each follow-up action was vetted by scientists on Earth, a single raster would take days or months to complete, wasting the rover's valuable time.

Instead, instruments like PIXL can benefit greatly from improved adaptive sampling methods like those proposed in our work. Adaptive sampling would allow the rover to analyze measurements as they are collected and determine the appropriate follow-up operation, greatly reducing the need for scientist oversight. Furthermore, the models generated by our approach as the rover collects samples could be used at future sample sites, reducing the number of redundant samples collected and improving overall sample diversity and instrument efficiency.

1.2 Contributions

This thesis will demonstrate techniques in adaptive sampling that allow a rover's understanding of its terrain to grow over time. We propose two probabilistic approaches that model an unknown number of latent classes in order to approximate both scene structure and an estimate of sampling reward. We iteratively sample from the scene using the low-resolution observations as a prior over possible measurements, choosing the sampling points most likely to improve either our models or the diversity of our collected samples. This process is sensor-agnostic, although elements of the approach are able to handle specific noise models or assumptions about data collected by a specific sensor. We make the following key observations:

Most scenes are made up of a small number of materials

Under this assumption, we propose that a small number of samples can be used to reconstruct nearly all of a scene. This assumption holds in both man-made and natural scenes, and our

approaches exploit this by iteratively sampling from areas dissimilar to our existing models, then growing the models dynamically over time.

Measurements from different sensors are correlated

In our outlined scenarios the use of low-resolution visual or spectral information is correlated with high-resolution spectra. In each of these cases there is overlap between the wavelengths of the data, leading to correlations even without explicit alignment or calibration.

Models persist between scenes

When observing natural terrain or objects, it is likely that materials and object characteristics will remain the same within a local region. This means any models built during one sampling operation can potentially be used as a prior in later sampling observations. Furthermore, informative prior information, such as multispectral maps, can be used to update the likelihood that our models do persist between specific scenes.

We develop two approaches that improve over existing methods in adaptive sampling. First, we use Dirichlet processes in model generation. Dirichlet processes estimate the parameters of classification models, such as the number of classes in the model. Dirichlet processes are a principled way to grow models dynamically over time. When applied to adaptive sampling they allow models to collect measurements, then decide if the information observed provides enough evidence to grow the model. This removes any constraints caused by a predefined model size. We show that Dirichlet models extend extremely well to the domain of microscopic sampling, when data is provided one point at a time and an immediate decision must be made about whether or not to sample that point. In these cases there is insufficient initial data to build more complicated models, and an extensible model like that shown here is able to adapt quickly and efficiently.

Second, we apply Bayesian optimization techniques for reward estimation. Bayesian optimization estimates the maximum of an unknown underlying function. We leverage work from this domain to estimate the diversity of new points, given our existing sample set. We differ from existing Bayesian optimization approaches in that our reward function, the underlying function we wish to optimize, changes at each iteration, but we find that these optimization approaches still offer competitive results.

Our proposed approaches have three advantages over related works in image analysis and autonomous science.

First, they are independent of spatial constraints, meaning they are not contingent on the assumption that nearby points are necessarily similar. Spatial constraints, while applicable to some domains, are often invalid in our scenarios. For example, the PIXL instrument can detect very small amounts of minerals that might only be within a single point measurement. If those minerals matched a known interesting spectral signature, we would want to sample, regardless of the local spatial context. Alternative methods, such as clustering approaches, would likely smooth this signal and be more likely to consider it noise with regard to its spatial context, rather than an interesting find.

Second, our approaches are entirely data-dependent and do not need to handle inter-instrument calibration. In many approaches that link multiple data products from different sources, an alignment step is necessary to provide good results. For example, when using orbital multispectral data and rover-based Visible to Shortwave Infrared (VSWIR) spectra, the products are typically aligned using a technique such as an empirical line approach [Gao et al., 2009]. This attempts to minimize differences in the products due to atmospheric interference, sensor calibration, and illumination, but is only an estimate on the true alignment parameters, calculated over a subset of recently-collected data. Instead, we observe that there is a correlation between our samples and their estimates, but we do not rely on any alignment step to detect this correlation.

Finally, our approaches work well in the noise that exists in these sampling scenarios and the extreme sparsity of the high-dimensional feature space. If the objective is to find maximally diverse sites, approaches must be able to model both the common and uncommon. This is a difficult problem when the observable data is noisy, as the true signal is often difficult to discern from points that appear interesting only due to noise. Our approaches are better able to adapt to these noise models and find interesting points within sparse and unmodeled regions of the feature space.

1.3 Overview of Dissertation

Adaptive sampling is a concept applicable to all scales of rover operations, from the microscopic to the orbital. In this work we will demonstrate techniques in adaptive sampling that perform well across all scales. We evaluate them on a number of experiments, designed to be similar to those performed during the day-to-day operations of a planetary exploration rover.

We begin with a discussion of our methods in more detail in Chapter 3. This chapter provides the mathematical background and theoretical context of our contributions.

Chapter 4 discusses published work in the field related to this thesis. Adaptive sampling is a field that has strong ties to multiple domains, such as active learning and information theory, and their contributions and similarities are discussed here.

Chapter 5 begins our series of experiments with a general case in which a rover views a noisy contextual dataset and must select points one at a time, with the goal of maximizing either the diversity of the dataset or the dataset's ability to reconstruct the scene, a measure of representativeness. We consider a large number of approaches, and give a detailed look at our contributions, comparison methods, and what pre-processing and implementation details result in the highest performance of each method.

Chapter 6 considers an application of our methods to data collected by the PIXL instrument. In this scenario, the rover no longer is able to fully observe the contextual data, but is instead provided one noisy sample at a time, as they are collected by the instrument. The rover must then immediately decide whether or not to follow up and collect a more detailed measurement at that location.

Chapter 7 considers an orbital path planning scenario, in which a rover plans a series of stops at promising measurement locations, the reward of which is estimated from orbital data. We compare

our contributions to state-of-the-art approaches in orbital path planning and sample site selection.

Finally, Chapter 8 concludes this thesis with a discussion of the contributions of this work, future avenues for research, and the implications of and need for adaptive sampling methods on both planetary and terrestrial robotics platforms.

Chapter 2

Motivations for Adaptive Sampling in Planetary Robotics

Today's remote exploration robots are well equipped to perform sampling tasks on command, but only have limited ability to autonomously select targets and analyze collected measurements. Adaptive sampling techniques can change and improve this, giving rovers the capability to understand their environment and the ability to recognize features of their environment that lead to informative sampling locations.

Improving adaptive sampling has two main benefits: increasing the productivity of exploration robots and enabling the use of instruments that would otherwise be prohibitively slow to operate in a planetary setting. Productivity is increased by enabling the rover to recognize diversity or select characteristic samples, then target these locations for measurements. This targeting and ability to understand its environment are then crucial components in enabling the use of advanced instruments, such as microscopic samplers that would be otherwise be prohibitively slow if operated with constant human input. These instruments require adaptive sampling techniques to select informative samples.

Finally, adaptive sampling techniques improve how exploration robots use their limited time during a mission. Time is a crucial resource; robots used in planetary missions only have a fixed number of sols before any number of mechanical, environmental, or resource problems can arise, limiting the rover's sampling ability. As a result, rovers are often directed by detailed plans developed by remote operations staff, such as the Science Operations Working Group providing daily science goals to the Mars Exploration Rovers [Mishkin et al., 2006].

These plans limit the time available for specific instruments or sampling, requiring either explicit targeting by the science team, or adaptive sampling methods to direct the instruments autonomously. Explicit targeting is appropriate in some scenarios, but new instruments, such as Planetary Instrument for X-Ray Lithochemistry (PIXL), must determine, at the time a measurement is collected, whether or not to follow up with more detailed measurements at that location. Back-and-forth communication between the instrument and a remote science team is not possible due to

communications delays, so adaptive sampling methods are necessary to improve the productivity of the instrument.

This chapter explores the larger context of this dissertation by examining two missions in planetary and exploration robotics where adaptive sampling methods are used. The first is the Mars 2020 mission. One of the objectives of this mission is to characterize the habitability of test sites, then search for evidence of past life on Mars by analyzing materials that have high biosignature preservation potential. Adaptive sampling methods in this mission are concerned with improving the information content of measurements collected by the rover's instruments, such as PIXL.

The second mission discussed is the Life in the Atacama (LITA) project [Wettergreen et al., 2013], in which an exploration rover traversed regions of the Atacama Desert in Chile. The goal of this project was to evaluate techniques in rover autonomy, characterize habitats within the desert, and search for subsurface signs of life. Our prior work in adaptive sampling was used on this mission to detect targets in navcam imagery and plan paths for the robot using orbital imagery. These methods improved the autonomy of the robot and increase the amount of science-relevant information returned to a remote science team, but in places fall short of their full potential.

2.1 Habitability Characterization by the Mars 2020 Mission

The Mars 2020 mission is an upcoming mission by NASA that is equipping a rover, similar in design to the Curiosity rover, with instruments designed to characterize the habitability and analyze astrobiological characteristics of the Martian terrain. The mission will also collect samples at a number of different locations, for a possible future return to Earth for analysis.

The goals of this mission highlight the importance of two of the evaluation metrics of this dissertation: diversity and representative sampling. The exact goals of the mission are as follows (emphasis added in italics) [NASA, 2016d]:

A *Characterize* the processes that formed and modified the geologic record within a field exploration area on Mars selected for evidence of an astrobiologically relevant ancient environment and geologic diversity

B Perform the following astrobiologically-relevant investigations on the geologic materials at the landing site:

- Determine the habitability of an ancient environment.
- For ancient environments interpreted to have been habitable, search for materials with high biosignature preservation potential.
- Search for potential evidence of past life using the observations regarding habitability and preservation as a guide.

C Assemble a returnable cache of samples for possible future return to Earth.

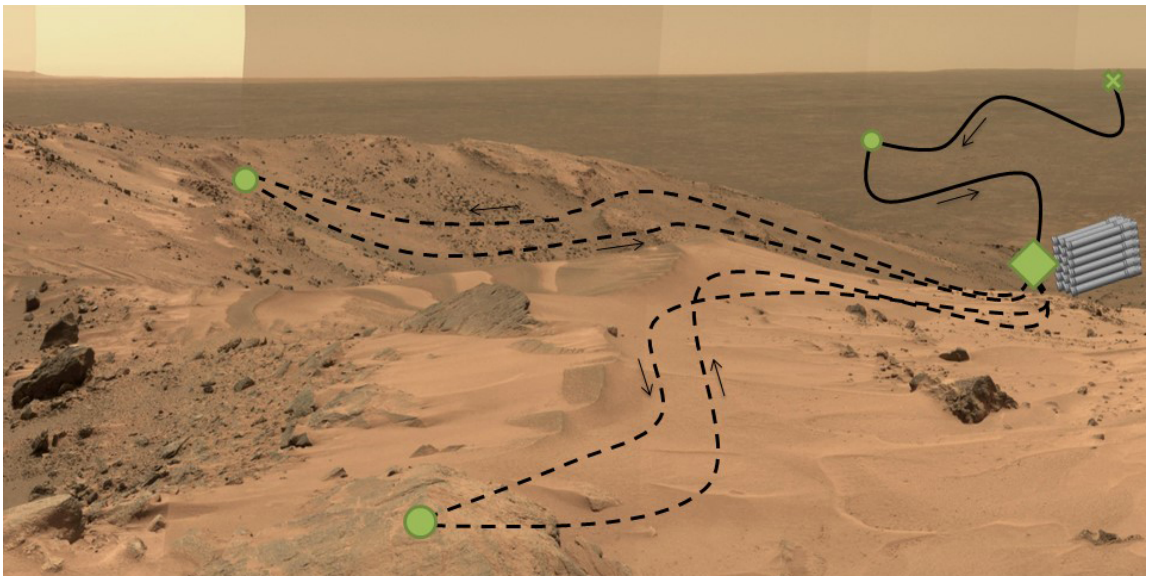


Figure 2.1: An example of the Mars 2020 sample caching concept, in which a rover lands at the position marked with 'x', then collects samples from diverse sites in the surrounding terrain and prepares them for sample return on a future mission [NASA, 2016a].

- Obtain samples that are scientifically selected; for which the field context is documented; that contain the most promising samples identified in Objective B; and, that represent the geologic *diversity* of the field site.
- Ensure compliance with future needs in the areas of planetary protection and engineering so that the cache could be returned in the future if NASA chooses to do so.

D Contribute to the preparation for human exploration of Mars by making significant progress towards filling at least one major Strategic Knowledge Gap. The highest priority SKG measurements that are synergistic with Mars 2020 science objectives and compatible with the mission concept are (in priority order):

- Demonstration of In-Situ Resource Utilization (ISRU) technologies to enable propellant and consumable oxygen production from the Martian atmosphere for future exploration missions.
- Characterization of atmospheric dust size and morphology to understands its effects on the operation of surface systems and human health.
- Surface weather measurements to validate global atmospheric models.

One main theme in these goals is the characterization of the Martian environment. Characterization here mirrors our objective of *representative sampling*, or the ability to collect measurements

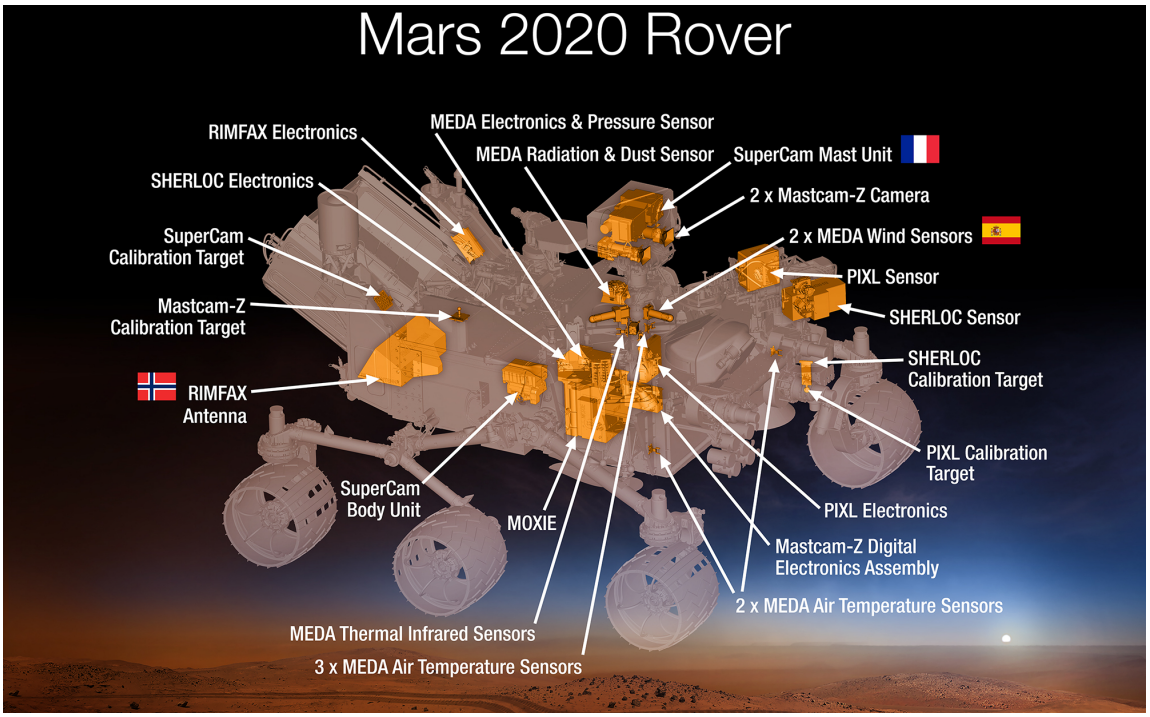


Figure 2.2: An overview of the instruments on the Mars 2020 mission [NASA, 2016b]. SHERLOC, SuperCam, and PIXL all have direct applications to adaptive sampling.

and samples that describe the environment. This is achieved by recognizing the constituent components of a scene, be that a small region sampled by a microscopic instrument, a field site, or a larger region of terrain, then detecting and collecting measurements from the exemplars within this scene. Adaptive sampling methods improve the rover's ability to determine these exemplars, then recognize and target them for more directed measurements.

The Mars 2020 mission is also interested in the search for diverse samples that might be returned to Earth. In this mission, this takes the form of finding sampling sites that offer diversity, in order to increase the variety of samples that are returned to Earth. These samples are likely to be a mixture of characteristic and diverse materials, representative of the region, but dissimilar or materials with high biosignature preservation potential.

These mission objectives motivate our own goals of building adaptive sampling methods that select diverse and representative points. By improving a rover's ability to estimate diversity or representativeness from a contextual observation, we pave the way for increased rover autonomy in future missions.

2.1.1 Applications of Adaptive Sampling on the Mars 2020 Mission

In addition to the overall goals of the Mars 2020 mission, the instruments on the rover also motivate improvements in adaptive sampling.

Figure 2.2 shows an illustration of the instruments on the Mars 2020 rover. Of these instruments, five are directly applicable to adaptive sampling or could be improved or made more autonomous by using adaptive sampling methods: Mastcam-Z, SuperCam, PIXL, Scanning Habitable Environments with Raman and Luminescence for Organics and Chemicals (SHERLOC), Radar Imager for Mars' Subsurface Experiment (RIMFAX).

Mastcam-Z is a mast-mounted camera system that collects imagery for the mission. This imagery can be searched for contextual cues to find good sampling locations, like unusual rock formations, veins, or locations to sample that are representative of the surrounding terrain.

SuperCam builds upon the success of the ChemCam instrument on the Curiosity rover, and uses Laser-Induced Breakdown Spectroscopy technology to determine the mineralogy of a target from range. Additionally, SuperCam can perform Raman spectroscopy, Time-Resolved Fluorescence spectroscopy, and Visible and InfraRed (VISIR) reflectance spectroscopy to analyze samples. This gives it a powerful toolkit to collect samples, and adaptive sampling techniques could be used to target the instrument or drive the use of multiple spectrometer readings if a first reading is estimated to have science-relevant information.

PIXL and SHERLOC are instruments that collect measurements on the microscopic scale. PIXL is an X-ray fluorescence spectrometer, and SHERLOC is a Deep UV resonance Raman and fluorescence spectrometer. Both of these instruments operate by collecting a raster of a small surface. Data is observed immediately, but the instruments can elect to observe the sample with a longer integration time to increase the signal-to-noise ratio. Adaptive sampling techniques are used here to evaluate the quick observation, then determine when to collect a long-duration spectrum. The adaptive sampling techniques used by PIXL are evaluated in Section 6, Adaptively Sampling with Limited Context, and data collected by PIXL is used in the experiments.

RIMFAX is a ground penetrating radar used to produce stratigraphic information about the subsurface of Mars. Adaptive sampling methods applied to RIMFAX could direct the path of the rover to follow interesting or promising readings. In prior work in conjunction with the Mojave Volatiles Prospector mission, we have evaluated a similar scenario in which a neutron spectrometer quantified subsurface hydrogen along the path of the rover [Foil et al., 2016]. We demonstrated that these readings could be correlated with orbital images of the scene, and improved with models of the physical processes that govern water transport in the desert, to predict subsurface hydrogen at unvisited areas of the terrain. In a similar manner, RIMFAX could make these predictions and use them as a contextual prior to drive adaptive sampling methods.

Out of the seven main instruments on the Mars 2020 rover, five of them already use adaptive sampling methods or could be directly applicable to adaptive sampling techniques. This demonstrates a clear need for development and study of adaptive sampling techniques.

2.1.2 Autonomous Path Planning in Future Planetary Missions

The goals of the Mars 2020 mission, specifically the goal of collecting a set of diverse samples across multiple locations, shows a possible future application of adaptive path planning methods.

Given orbital imagery as context and the prior set of collected samples, adaptive sampling methods are well-suited to estimating the diversity of future sampling sites. The rover compares the orbital measurements at sampled locations with the measurements of the samples collected at those locations, then chooses new sample sites based on the data collected at prior stops. The samples collected during the Mars 2020 mission are of such high importance that they will be hand-selected by scientists, yet future missions in which a rover is given more autonomy could use adaptive sampling in this manner.

This scenario is explored in Chapter 7, Adaptive Path Planning, in which we evaluate different adaptive sampling techniques on their ability to generate paths of sampling stops that maximize the reward of the set of samples collected.

2.2 Life in the Atacama

The Life in the Atacama project (LITA), was a recent investigation into planetary robotics methods for rover autonomy, desert habitat characterization, and techniques to search for subsurface signs of life. The mission consisted of two field tests in the Atacama Desert, Chile, in which the rover, Zoë, drove approximately 100 kilometers, collecting thousands of reflectance spectra, dozens of measurements from a Raman spectrometer, and multiple soil samples for laboratory analysis.

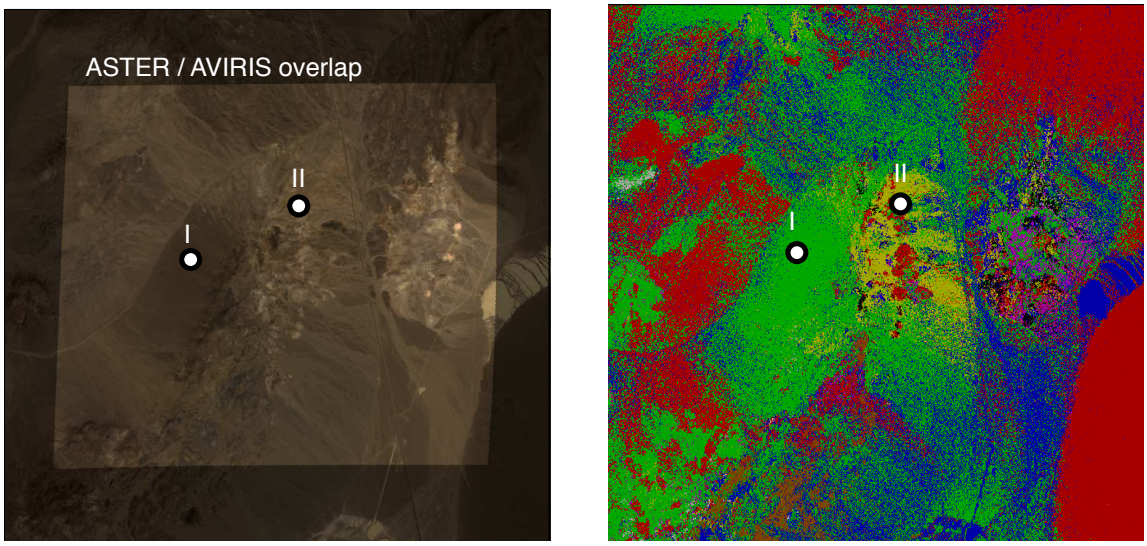
During this project, we evaluated adaptive sampling methods, aimed at providing a rover with the autonomy needed to identify targets or regions of interest, then collect samples that contained information relevant to the science goals of the mission.

We tested two main techniques: path planning using orbital data and target selection using image analysis. We found that both methods improved the ability of the rover to collect informative samples, yet those two approaches were inadequate, motivating further work in adaptive sampling research.

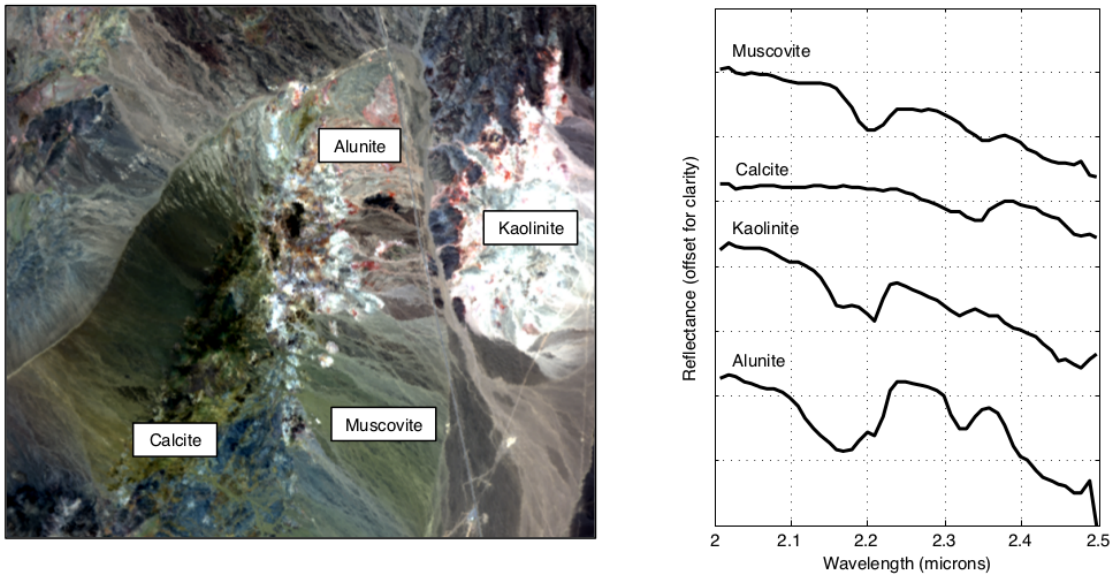
2.2.1 Spatio-Spectral Exploration

Our work on spatio-spectral exploration [Thompson et al., 2015b], commonly called **SpectroMapper**, is an application of adaptive sampling that was used in the LITA mission to plan sampling paths for Zoë. SpectroMapper uses a library of orbital and rover-collected spectra to build up a representative collection of samples, ultimately trying to collect the subset of sampled points that can best reconstruct the original map of the terrain.

The goal of the spatio-spectral exploration approach is to generate the path that collects the most diverse set of spectra possible between a start and end point in a traverse, all limited by an exploration budget. To do this, we first think of the orbital image as a mixture of *endmember spectra* [Keshava and Mustard, 2002]. Endmembers are characteristic spectra, and, in the absence



(a) Cuprite, NV test site and associated endmember classes



(b) Geologic endmembers at a nearby location

Figure 2.3: Top left: an aerial view of Cuprite, Nevada. Top right: the same view colored according to it's dominant spectral class, or endmember. Each color represents a different dominant endmember at that location [Thompson et al., 2015b]. Bottom left: a nearby location at Cuprite, NV with more diversity. Bottom right: the geologic classes present in this scene. Note that endmembers can, but do not always, correspond to geologic classes.

of error, one could collect all the endmember spectra in a scene and fully reconstruct every point. This is similar to Principal Component Analysis or other dimensionality-reduction techniques. For an example see Figure 2.3. Endmembers represent pure pixels within the overall scene, and often correspond with the recognizable macroscopic objects that we might observe ourselves, such as lakes, sand, or forests.

In the SpectroMapper case, we want to collect the library of spectra that best reconstructs the entire orbital image. Furthermore, we expect the endmember library Y to be compact and that it can be used to reconstruct any measurement $\mathbf{x}$ with a linear mixing model using mixing fractions ϕ, such that $\mathbf{x} = \phi Y$.

Formally, we aim to collect the set of in-situ measurements that best unmix, or reconstruct, the orbital spectra. We do this by evaluating a set of candidate measurement sites in the orbital imagery, $\mathcal{L}$, and selecting the subset of sampling locations $B = \{b : b \in \mathcal{L}\}$ such that the spectroscopic library generated by samples collected at B, Y_B, minimizes the expected reconstruction error of the entire remote image:

$$R(B) = E\left[\sum_{\mathbf{x}\in X} \min_{\phi} ||Y_B\phi - \mathbf{x}||_2\right] \tag{2.1}$$

for a non-negative set of mixing fractions ϕ and subject to some overall budget $C(B) < \beta$. Non-negativity has a physical basis and interpretation here; the spectrum used in the reconstruction here correspond to measurements of physical objects, and, as such, cannot be negative.

As we have not yet collected spectra at each location, we approximate them using remote data at each location, X_B. The equation then becomes

$$R(B) = \sum_{\mathbf{x}\in X} \min_{\phi} ||X_B\phi - \mathbf{x}||_2 \tag{2.2}$$

As the rover collects spectra, they are aligned with the orbital data to reduce any atmospheric, instrument, or lighting effects that might cause misalignment between both sets of spectra. This alignment is done using a modified empirical line approach [Gao et al., 2009] which computes a linear offset between the common wavelengths of both rover and orbital instruments, averaged across the entire collected dataset.

Once aligned, the collected spectra, Z_A, are added to the onboard library of spectra and are used as more accurate data for endmember calculation. The equation then becomes

$$R(B|A) = \sum_{\mathbf{x}\in X} \min_{\phi} ||[Z_A\ X_B]\phi - \mathbf{x}||_2 \tag{2.3}$$

again for $\phi > 0$ and respective of a total budget $C(A) + C(B) < \beta$.

Overall, this results in the following behavior: given a target location, the rover evaluates points on a regular grid between the current location and goal. The remote spectra at each location are

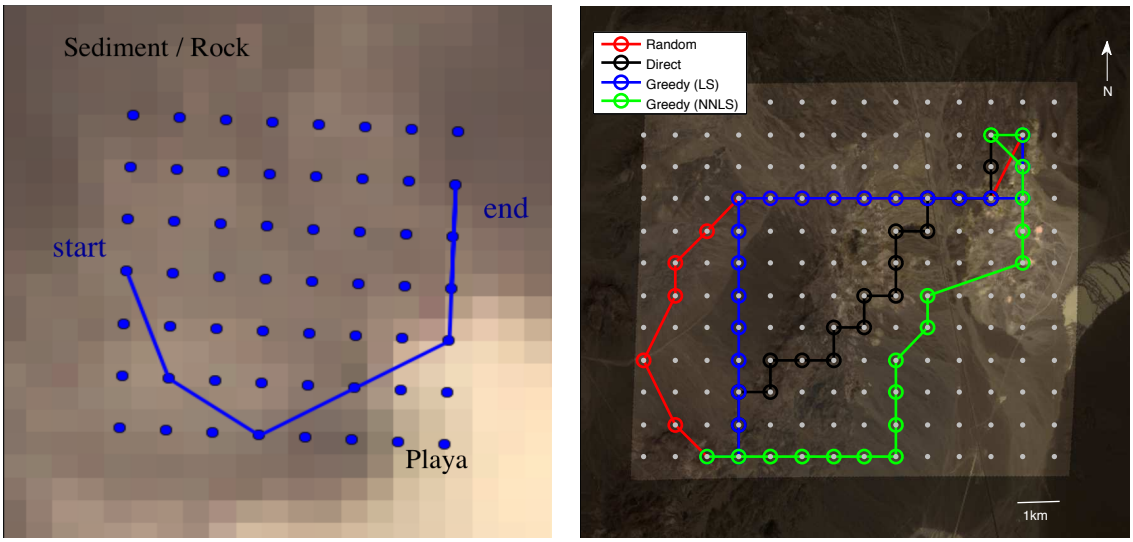


Figure 2.4: Two demonstrations of the spatio-spectral exploration approach. In each case the rover is given a goal and must find the sampling points within the grid that provide the lowest estimated reconstruction error of the entire scene, constrained by a budget on total travel distance. Left: A path planned across the north end of a playa during Atacama field tests in 2013. Right: a set of paths planned in simulation in which our approach using two least-squares solvers is compared against random and direct approaches [Thompson et al., 2015b].

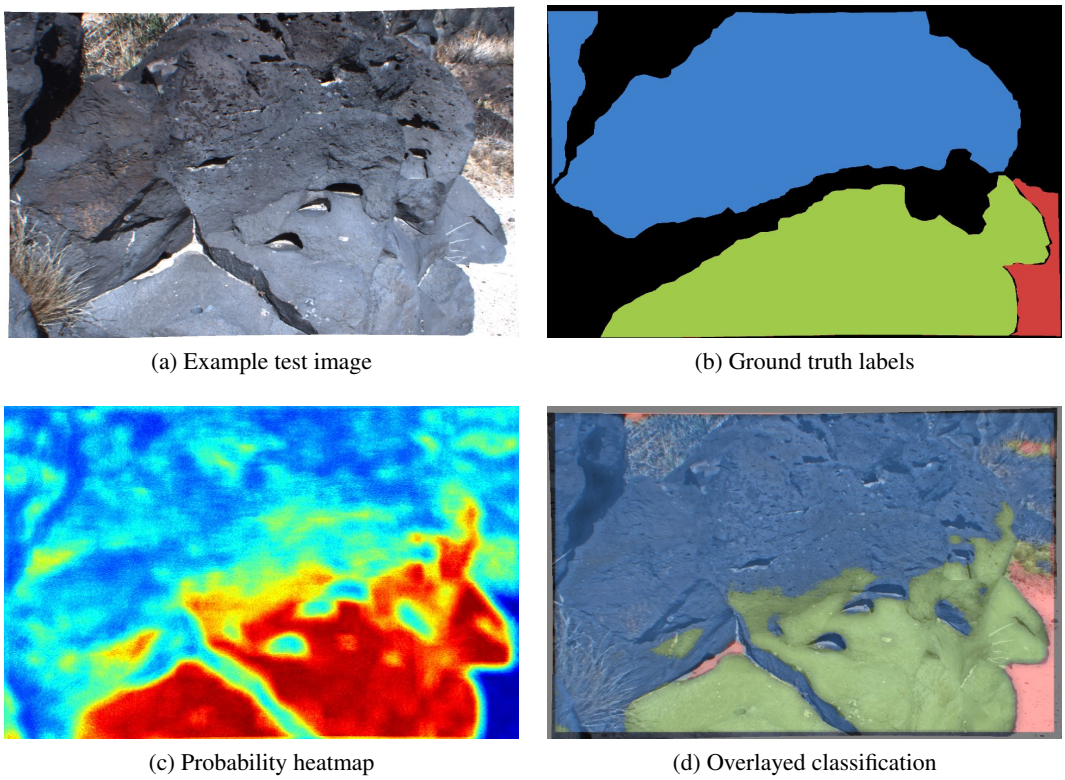
then compared to the onboard library of collected spectra. At each candidate measurement point, the error is calculated between the library's reconstruction of that point and the true data at the point. The rover then plots a path through the candidate points, planning to stop and sample at the points with the highest error. Upon sampling any point, the rover-collected data is aligned with orbital data, endmembers and the residual errors at each point are recalculated, and a new path is planned. This repeats until the rover has reached the goal or has run out of budget, at which point it heads to the goal and completes the traverse.

Through experimentation we demonstrated that this approach outperforms existing alternatives, resulting in more efficient exploration. Furthermore, by altering the objective function to focus on non-negative, sparse endmember reconstructions, as one would typically find in a real-world environment, we saw performance improvements. We further tested this approach in the field, and sampled a much more spectrally diverse set of locations than we otherwise would have sampled.

SpectroMapper is used as a baseline approach in the adaptive path planning experiment in Chapter 7.

2.2.2 TextureCam

The LITA project also used intelligent image processing methods to detect targets of interest on command. Image understanding is an important part of intelligent robotics and is often overlooked



(a) Example test image

(b) Ground truth labels

(c) Probability heatmap

(d) Overlayed classification

Figure 2.5: Image classification using the TextureCam algorithm. Top left: a candidate image for classification. Top right: a geologist's ground truth labeling. Blue is vesicular basalt and green is smooth basalt. Bottom left: probabilistic heatmap showing predictions of the smooth basalt class. Bottom right: output classification predictions made by TextureCam [Foil et al., 2013].

or simplified in planetary missions due to the high computational cost of advanced image processing algorithms. TextureCam, the approach used on Zoë, is a project that aims to provide efficient and robust image classification for planetary rovers using parallelizable random forest-based classification routines implemented on an FPGA.

TextureCam is a supervised classification routine that operates by generating a large number of decision trees T trained on random subsets of pre-classified training data. Nodes in each tree are simple functions operations on a single pixel or pairs of pixels within a small window around a candidate classification point. The leaves of each tree, l, contain the probability $P(c_i|l)$ of a given class $\{c_i \in C\}$ reaching that node given the set of nodes used to reach that leaf. The classification for a given pixel in the scene is thus the class with the largest average over all leaf node probabilities:

$$\arg\max_{c} \frac{1}{T} \sum_{t=1}^{T} P(c|l_t) \tag{2.4}$$

Each decision tree is independent, and thus the entire forest is highly parallelizable, making it ideal for a hardware implementation on an FPGA device. For more details on the TextureCam algorithm and approach, please see [Foil et al., 2013, Wagstaff et al., 2013]. An example is shown in Figure 2.5.

TextureCam can also be for target detection and followup. In this framework, TextureCam classifies navcam images captured by the rover, finding high-value targets like rocks. The largest target in an area near the rover is selected and a spectral raster follows up centered on this point. An example of the process is shown in Figure 2.6

Our work on TextureCam focused on extending the functionality to characterize regions of high confidence classification regions and novel regions [Foil et al., 2013]. To do this, we implemented an abstention approach that computed the expected loss of choosing to classify a pixel or abstain from classification. Correct classifications are given no cost, misclassifications are given a per-class cost of β_c, and abstentions are given an abstention cost $\{\alpha : 0 \leq \alpha \leq \min_c \beta_c\}$. Given n total classes, the expected loss of choosing action a is:

$$E[L|x,a] = \frac{1}{n}\sum_c L(c,a)P(C=c|x) \tag{2.5}$$

$$\text{where } \mathrm{L(c,a)} = \begin{cases} 0 \text{ if } \mathrm{a=c} \\ \beta_c \text{ if } \mathrm{a} \neq \{\mathrm{c, abstain}\} \\ \alpha \text{ if } \mathrm{a = abstain} \end{cases} \tag{2.6}$$

It is optimal to choose the action that is associated with the smallest expected loss. Extrapolating to m trees $(t_1, ..., t_m)$, the optimal action is found by averaging the expected loss across all trees:

$$\arg\min_a \; E[L|x,a] = \frac{1}{n}\sum_c L(c,a)\frac{1}{m}\sum_t p_t(C=c|x) \tag{2.7}$$

This formulation gives both a framework for finding both high confidence classification regions. High confidence regions are useful for scenarios in which representative samples with the highest probability of success are desired, such as a probe landing on an environmentally-hostile planet and only having one sampling attempt.

Alternatively, we may be interested in low confidence regions, as they are potentially indicative of novel signatures or unknown classes. Using the formulation given above, we extended the classification of high confidence regions to further model novelty within the scene. This was done by noting that the confidently classified regions in the image had very low classifications in testing. By working under the assumption that these regions had low classification error, on average, we could then model the color and textural distributions of each confident class, creating a Gaussian mixture model (GMM) representing our known classes. We then compared regions of low classification confidence, m, to our GMM, selecting regions for follow-up that had the greatest minimal Mahalanobis distance between itself and any known class:

$$x_{n+1} = \arg\max_{x_i \in X} \; \arg\min_{c \in C} \sqrt{(x_i - \mu_c)^T S_c^{-1} (x_i - \mu_c)} \tag{2.8}$$

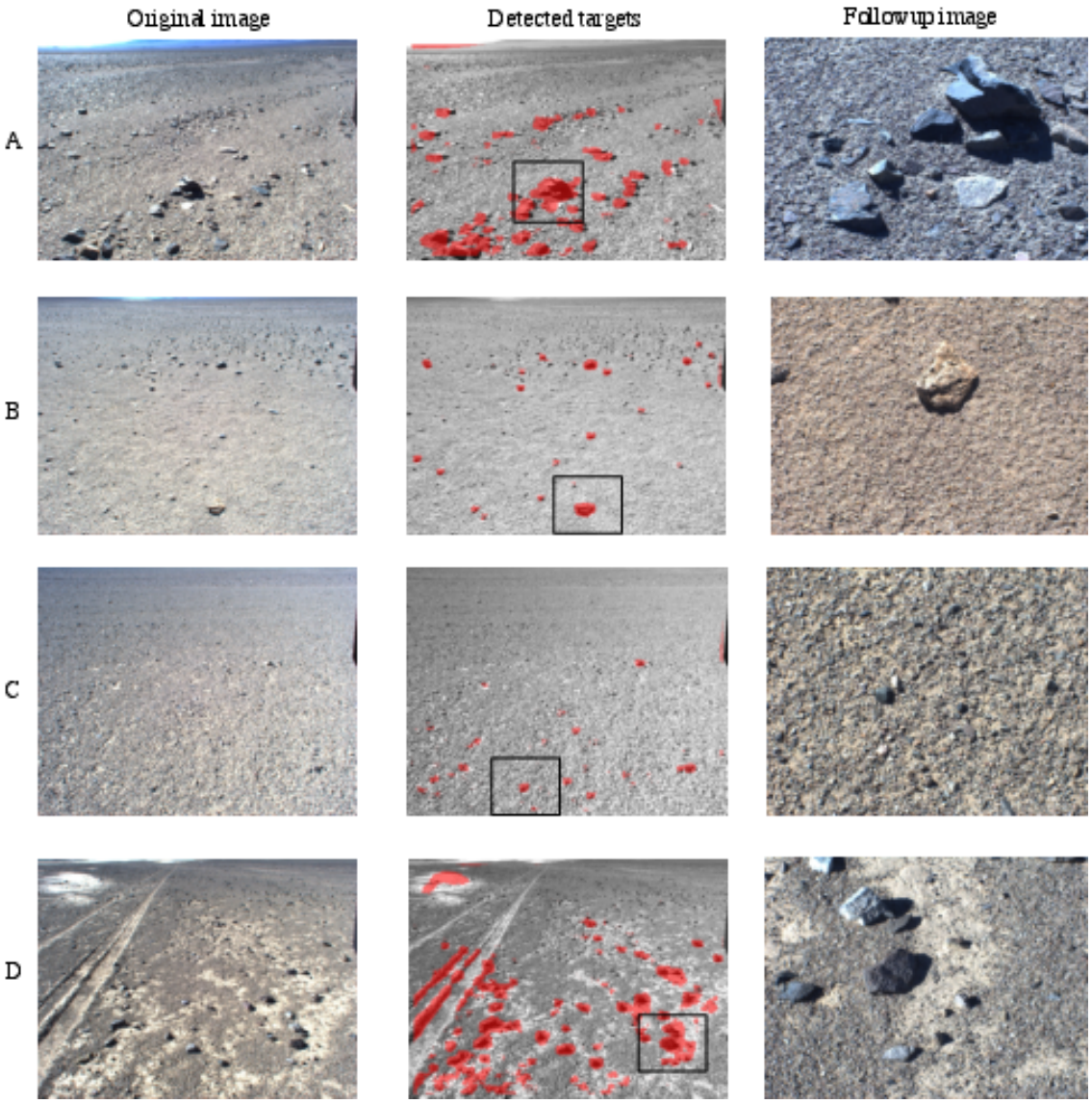


Figure 2.6: An example of the targeting approach using TextureCam results showing four examples in rows A through D. Left column: the original navcam images collected by the rover. Center column: classification output of the TextureCam algorithm. Right column: images of follow-up targets at the center of a spectral raster [Wettergreen et al., 2013].

where μ_c and S_c are parameters of the multivariate Gaussian model for class c.

2.2.3 Observing Context: The Missing Link

The SpectroMapper and TextureCam approaches enable the rover to plan and sample at the orbital and rover scales. In the LITA tests, these approaches were used together; paths were planned by the SpectroMapper, then, upon reaching a target location, TextureCam would observe the terrain and choose sampling targets that were likely to contain scientifically-relevant data.

However, there are major gaps in this approach. Due to the spatial resolution of orbital data, the rover is directed to a very broad location, typically 15 or 30 meters square. The rover drives to the target location, then stops and samples directly in front of where it stopped.

This often leads to a disconnect between the material targeted by the SpectroMapper and the sample collected by the rover. If the ground cover at the sampling location is not uniform, then there is no guarantee that the target material is even observable from the point at which the rover stopped.

Furthermore, no attempt is made to analyze the data collected between the sampling locations determined by SpectroMapper. This data, primarily color navcam images, offers a huge amount of contextual information that can be used to direct the rover to informative sampling locations. Imagery contains color, textural, and structural information that can be used to identify interesting rocks or promising features in the terrain. Additionally, it can be calibrated to the visible channels of orbital images, giving a rough prior on the feature of interest targeted by the rover when a given sample site is chosen. Anecdotally, there were multiple times when the rover was directed to a location at which there was one distinct terrain feature that stood out from the surrounding region, like a distinct ground cover or a small playa, yet the rover failed to sample the distinct region and instead stopped at an arbitrary point with the goal area and sampled directly in front of the rover. In these cases, if a rough estimate of the visible characteristics of the region can be estimated from the orbital imagery, the rover can monitor navcam imagery and make a more informed decision about where to stop and sample. If nothing else, the rover should be able to collect a measurement, evaluate it against what it expected to find, and determine whether or not the correct target was sampled or if it should scan the surrounding area for more promising sampling locations.

2.3 Improving Rover Capabilities with Adaptive Sampling

These two missions illustrate the growing need for adaptive sampling methods. Adaptive sampling is already being applied in planetary missions, yet there are still many gaps in the exploration process which could be filled with improved adaptive sampling methods.

These missions also show the need for both diverse and representative sampling. Diverse sampling recognizes interesting, unknown, or under-represented signatures. Representative sampling characterizes regions, providing detailed measurements for model or instrument validation and greater understanding of the dominant constituents of a scene. Both are important in varying degrees, depending on mission goals, and they form the basis for our quantitative analysis.

Chapter 3

Approaches to Adaptive Sampling from Context

In this section we detail the methods that will be used for sample selection and path planning and will demonstrate improvements over existing sampling and planning methodologies. The main goals of our work here are to find approaches with some or all of the following characteristics:

A The ability to take advantage of underlying trends in the data, like abundances in a single element across all candidate samples.

B The ability to distinguish noise in the data from meaningful changes in geophysical variables.

C Flexibility in sampling from an unknown number of underlying classes.

This work proposes two approaches with qualities that satisfy these characteristics. First, in Section 3.1, we introduce a mixture model-based approach and its extension to use a Dirichlet process prior to generate new classes and predict misclassifications. These mixtures allow the model to characterize different mineralogical classes, estimate the noise present in each class, and flexibly expand to create room for new classes as evidenced by the data.

Second, in Section 3.2, we introduce a method that estimates the reward of future samples using Gaussian process models, allowing us to use established Bayesian optimization techniques to select points that have a high expected reward. This approach is less susceptible to noise in the model, as expected rewards are typically estimated from a combination of multiple points with similar features, while also providing an actionable measure of model uncertainty that allows for the testing of a variety of different sampling metrics.

In each set of experiments we compare our approach against a number of different methods, detailed in Section 3.3.

3.1 Adaptive Gaussian Mixture Models

Our first method advances the work of Sillito et al. [Sillito and Fisher, 2007] and their incremental kernel density estimation. We consider the creation of a mixture model, similar to a Gaussian mixture model (GMM) [McLachlan and Basford, 1988].

GMMs are a method for clustering data, in which a fixed number, K, of Gaussian mixtures is created that best fits the given training data. Each mixture is characterized by a mean vector and covariance matrix, the parameters of which are typically learned through expectation maximization (EM) over the training data [Dempster et al., 1977]. New points are then assigned a class based on their similarity to each Gaussian mixture.

Sillito et al. consider the specific case in which a mixture model is incrementally generated and cannot be learned a priori using training data. In their case, they describe an approach for outlier detection using incrementally-learned density estimation. They initialize their approach by using kernel density estimation (KDE) [Rosenblatt, 1956], in which they estimate the probability of a new point by taking a uniformly-weighted combination of Gaussian kernels centered at all prior sample points. Each kernel has an identical covariance matrix Σ, so the probability of a new point, x_{n+1}, given the prior N sample points $X = \{x_1, ..., x_n\}$ and dimensionality d, is simply

$$P(x_{n+1}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \frac{1}{N} \sum_{n=1}^{N} e^{-\frac{1}{2}(x_{n+1}-x_n)^T \Sigma^{-1}(x_{n+1}-x_n)} \tag{3.1}$$

When new points are collected, they are added to the sample set:

$$X \rightarrow X \cup x_{n+1} \tag{3.2}$$

As mentioned above, the kernels have identical covariance matrices. To reduce computational costs, the covariance matrices are diagonal with equal variance in each channel: $\Sigma = I^d \cdot \sigma^2$. In order to estimate the value of σ, they recommend the "leave-one-out" method proposed by Duin [Duin, 1976], in which a number of values of σ are evaluated on the dataset, and the one that maximizes the likelihood of the dataset, σ^* is chosen:

$$LL(\sigma_i) = \sum_{n=1}^{N} \log \left(\frac{1}{(2\pi\sigma)^{\frac{d}{2}}} \frac{1}{N-1} \sum_{\forall x \neq x_n} e^{-\frac{1}{2\sigma^2}(x_n-x)^T(x_n-x)} \right) \tag{3.3}$$

$$\sigma^* = \arg\max_i LL(\sigma_i) \tag{3.4}$$

If the full data is available *a priori*, σ can simply be evaluated once over the entire range of data. If the data appears incrementally, σ can be re-evaluated periodically, typically by initializing it over a wide range of candidate values at the start of the experiment, then refining it over a smaller range as data is collected.

This process grows linearly with the number of sample points, and eventually becomes computationally burdensome to evaluate the density of new points. Instead, Sillito et al. note that

this formulation is equivalent to a Gaussian mixture Model (GMM) with N mixtures, and adapt a method proposed by Goldberger and Roweis [Goldberger and Roweis, 2005] to reduce the complexity of the model.

They do this by setting a maximum number of mixtures, N_{max}. If $N \leq N_{max}$, the model is created as described above. Once $N = N_{max}$, a GMM with N_{max} mixtures is created with the following parameters:

$$w_{1...N} = \frac{1}{N} \tag{3.5}$$

$$\mu_{1...N} = x_{1...N} \tag{3.6}$$

$$\Sigma_{1...N} = I^d \cdot \sigma^* \tag{3.7}$$

where w_i, μ_i, and Σ_i are the weights, means, and covariances of the ith mixture, respectively.

Once $N > N_{max}$, when the number of mixtures has reached its limit, the algorithm decides whether to merge the new data with an existing mixture, or merge two mixtures together and fill the now-empty spot with a mixture initialized with the new sample. In order to make this decision, a merging cost is calculated between all existing mixtures and a proposed new mixture with statistics generated from the new sample.

The cost of merging mixtures is defined by the Kullback-Leibler (KL) divergence:

$$\mathcal{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{3.8}$$

This encodes the expected information loss when an approximating distribution, Q, is substituted for a true distribution, P. For Gaussian mixtures, this is defined as

$$\mathcal{KL}(G_p, G_q) = \frac{1}{2} \left(\log \frac{|\Sigma_q|}{|\Sigma_p|} + \text{Tr}(\Sigma_q^{-1}\Sigma_p) + (\mu_p - \mu_q)\Sigma_q^{-1}(\mu_p - \mu_q)^T - d \right) \tag{3.9}$$

where d is again the dimensionality of the data. Using this allows us to calculate the cost of replacing two mixtures, G_i and G_j, by their merged counterpart, $G_{merge(i,j)}$, by calculating the weighted combination of their KL divergence:

$$c(G_i, G_j) = w_i \mathcal{KL}(G_i, G_{merge(i,j)}) + w_j \mathcal{KL}(G_j, G_{merge(i,j)}) \tag{3.10}$$

When the mixture model is initialized, a $N \times N$ cost matrix C is formed, where $C(i, j)$ is the cost of merging mixtures i and j. There is a one-off calculation cost of $\frac{N_{max}(N_{max}-1)}{2}$ to initialize the matrix. Then, as new points are sampled, the cost of merging the new point with existing mixtures is calculated using the above method. If the cost of merging the new point, c_{new}, is less than any value in the merge cost matrix C, the point is added to the most similar mixture and the mixture's weights and statistics are updated. If the cost of merging the new point, c_{new}, is higher than the minimal value in the cost matrix C, then the two mixtures with the lowest merge cost are merged, as selected by $\arg\min_{i,j} C(i, j), i \neq j$. The new point is then added to the mixture model in the now open spot, and the relevant statistics and cost matrix values are updated.

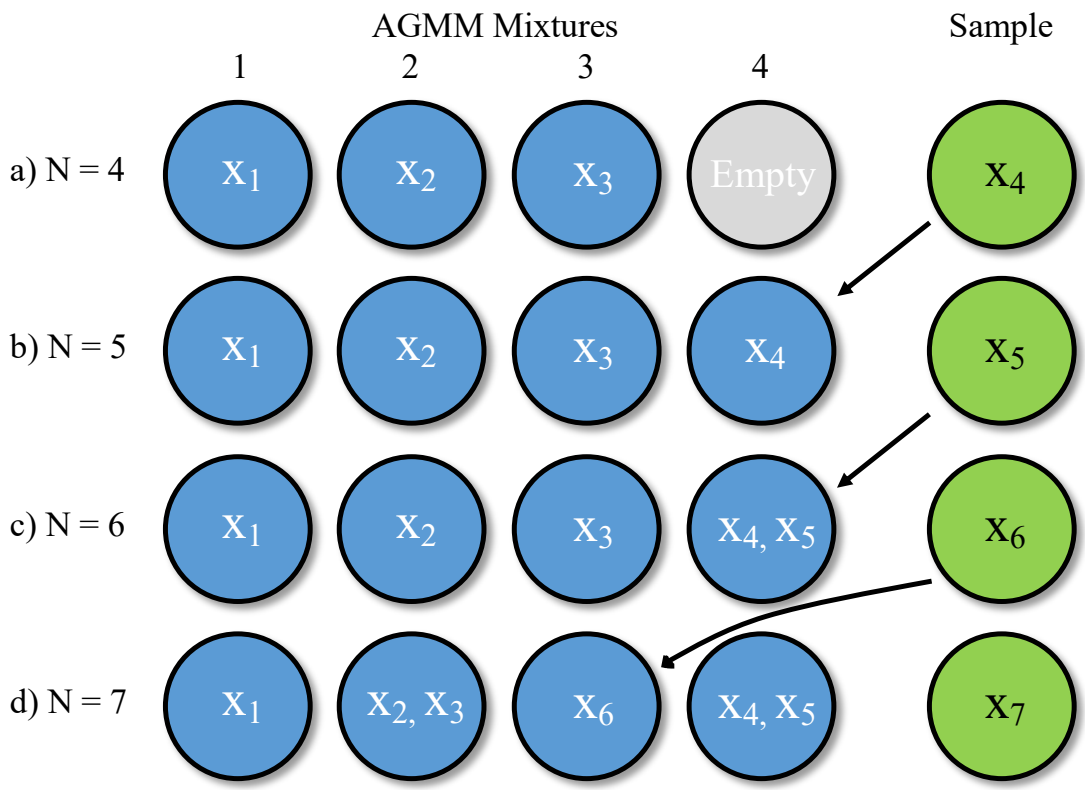


Figure 3.1: An example of the AGMM process at work. Arrows indicate to what mixture a sample was assigned prior to collecting the next data point. a) Sample x_4 is collected and assigned to the last open mixture. b) Sample x_5 is collected that has a lower cost of merging with mixture 4 than the cost of merging any existing clusters. c) Sample x_6 is collected that has a higher merge cost than the cost of combining mixtures 2 and 3. d) Mixtures 2 and 3 are merged together and x_6 is assigned to the empty position.

The merging operation of two mixtures, i and j, is done as follows:

$$w_{merge(i,j)} = w_i + w_j \tag{3.11}$$

$$\mu_{merge(i,j)} = \frac{w_i}{w_i + w_j}\mu_i + \frac{w_j}{w_i + w_j}\mu_j \tag{3.12}$$

$$\Sigma_{merge(i,j)} = \frac{w_i}{w_i + w_j}(\Sigma_i + (\mu_i - \mu_{merge(i,j)})(\mu_i - \mu_{merge(i,j)})^T) \tag{3.13}$$

$$+\frac{w_j}{w_i + w_j}(\Sigma_j + (\mu_j - \mu_{merge(i,j)})(\mu_j - \mu_{merge(i,j)})^T) \tag{3.14}$$

We refer to this as our Adaptive GMM approach (**AGMM**). Its formulation is computationally cheap, a strong advantage to approaches targeted at planetary robotics scenarios, and is flexible

enough to work well at finding and sampling diverse points in our dataset. However, the traditional Gaussian mixture model approach assumes prior knowledge of the appropriate number of mixtures in the model. Too few mixtures leads to an over-generalization of the data, while too many mixtures can lead to over-fitting. This presents a difficult problem in our work, as the number of mineralogical classes is unknown beforehand, if there even are well-defined classes within the data.

There are many remedies to this model selection problem. Most notable are Akaike information criterion (AIC) [Akaike, 1974] and Bayesian information criterion (BIC) [Schwarz et al., 1978]. AIC is defined as

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{3.15}$$

where k are the number of free parameters in the system and $\hat{L}$ is the maximum value of the likelihood function of the model in question. Given a number of candidate models, the best model is the one with the minimum AIC score. AIC penalizes the number of free parameters, as increasing the number of model paramters often leads to over-fitting.

BIC is defined as

$$\text{BIC} = -2\ln(\hat{L}) + k\ln(n) \tag{3.16}$$

where $\hat{L}$ is again the maximum value of the likelihood function of the model, k are the number of free parameters, and n are the number of data points in the observed data. Similar to AIC, BIC penalizes the number of free parameters, weighted by the size of the data, in order to reduce model complexity.

In practice, each of these scoring criterion can be used to estimate an appropriate number of mixtures in our AGMM formulation. However, we instead focus on a novel method for the addition of classes that does not necessitate re-learning new models or adjusting the mixtures that have already been created.

We propose a method for adaptively increasing the number of mixtures in our model by using a Dirichlet process prior. This prior is over the number of mixtures in the model, and allows the model to be flexible in a principled and probabilistic manner. New new classes are added when a radically diverse point is discovered and points similar to the existing model are merged with existing classes. Over time this enables the model to have a number of classes more similar to the true underlying distribution of classes, while also merging similar points in a more appropriate manner.

In the following section we provide an overview of Dirichlet processes. In Section 3.1.2 we discuss the details of an approach for adding new classes in our AGMM using a Dirichlet process priors to improve classification. In Section 3.1.3 we discuss using a Dirichlet process prior to better estimate the probability of misclassification of candidate sample points.

3.1.1 Dirichlet Process Priors to Estimate Model Parameters

In many sampling and active learning approaches, classifiers and models are developed using observed data. In most classification-based approaches, such as Gaussian mixture models, the number

of classes is assumed to be known or set to a fixed value in advance. In some domains this is a valid assumption, but in this work the number of classes is unknown beforehand and may even grow over time.

To solve this problem, the use of Dirichlet distributions and Dirichlet processes was proposed as a way to formalize class assignment under uncertainty [Ferguson, 1973]. *Dirichlet distributions* are distributions over continuous values summing on one. They are a conjugate prior to multinomial distributions, meaning that if a data point is drawn from a multinomial distribution whose parameters are Dirichlet distributed, the posterior estimate of the multinomial is also Dirichlet distributed. This means that the estimate of the parameters of a multinomial can be successively refined and updated, one observation at a time.

Dirichlet processes are generalizations of these distributions that handle an infinite number of parameters for an unknown number of classes. They are often used in situations such as the adaptive sampling scenario described here, where both the parameters of each model and number of models is unknown.

Mathematical Background

Say we observe $X = \{x_1, ..., x_n\}$, where X are independent and identically distributed draws from a distribution F. In Bayesian modeling, one would typically place a prior over F, then calculate the posterior on F given the data X. This prior over distributions is typically given by a parametric family, but this assumption is limiting. Instead, a nonparametric approach, such as the Dirichlet process, can be used as a prior over distributions with wide support, typically the the space of all distributions.

To understand Dirichlet processes, first let us start with an explanation of the Dirichlet distribution. Consider building a Gaussian mixture model from the data X mentioned above. The corresponding model is

$$P(x_i|c_i = k, \Theta) \sim \mathcal{N}(\cdot|\theta_k) \tag{3.17}$$

$$P(c_i|\vec{\pi}) \sim \text{Multinomial}(\cdot|\vec{\pi}) \tag{3.18}$$

where $C = \{c_i\}_{i=1}^N$ indicates the class that each point of data belongs to, $\Theta = \{\theta_k\}_{k=1}^K$, $\theta_k = \{\mu_k, \Sigma_k\}$ are the parameters of the Gaussian distributions of each class, and $\vec{\pi} = \{\pi_k\}_{k=1}^K$ are the mixture weights.

A simple approach to estimate $\vec{\pi}$ and Θ might be to use Expectation Maximization (EM) [Dempster et al., 1977] to find the maximum likelihood of the model parameters:

$$\hat{\vec{\pi}}, \hat{\Theta} = \arg\max_{\vec{\pi}, \Theta} \log(P(X, C|\vec{\pi}, \Theta) \tag{3.19}$$

However, while EM is guaranteed to converge to a local maximum, it will not necessarily find the global maximum and requires a good initialization point. Furthermore, the likelihood of the data can be increased by increasing K, so there must be constraints on the model so $K = N$ is

not chosen. Typically this is done by constraining the model complexity, using something like the Bayesian information criterion:

$$BIC = -2\log(P(X, C|\vec{\pi}, \Theta)) + \nu_K \log(N) \tag{3.20}$$

where ν_K is the number of free parameters in a model with K densities [Schwarz et al., 1978].

However, we would like to avoid these problems and focus instead on a more Bayesian approach. Momentarily assume that we know the value of K here, the number of classes in our mixture. We can then place a prior over our model parameters:

$$\vec{\pi}|\alpha \sim \text{Dirichlet}(\frac{\alpha}{K}, ..., \frac{\alpha}{K}) \tag{3.21}$$

$$\Sigma_k \sim \text{Inverse-Wishart}_{\upsilon_0}(\Lambda_0^{-1}) \tag{3.22}$$

$$\mu_k \sim \mathcal{N}(\mu_0, \Sigma_k/\kappa_0) \tag{3.23}$$

The Dirichlet concentration parameter, α, encodes our belief about how uniform or skewed the weights $\vec{\pi}$ will be, acting as an inverse variance. A low value of α will lead the distribution to favor assigning points to only a handful of classes, or placing points in existing classes in the case where K is infinite. A high value of α will lead the distribution to assigning points to a large number of classes, or creating additional classes if K is infinite. The parameters to the Inverse-Wishart prior also encode our prior knowledge about the mixture distributions. The Dirichlet distribution and Inverse-Wishart priors are chosen because they are conjugate to the multinomial and normal distributions, respectively.

Under this model, we can compute the posterior distribution:

$$P(C, \Theta, \vec{\pi}, \alpha|X) \propto P(X|C, \Theta)P(\Theta|\mu_k, \Sigma_k)\prod_{i=1}^{N} P(c_i|\vec{\pi})P(\vec{\pi}|\alpha)P(\alpha) \tag{3.24}$$

Computing this analytically is impossible, but we can sample from it with Markov Chain Monte Carlo (MCMC) methods [Neal, 2007], explained in more detail in Section 3.1.1. Sampling from this distribution avoids the initialization and local optima problems from which EM suffers.

Now the only terms of Equation 3.24 that involve K are $P(c_i|\vec{\pi})$ and $P(\vec{\pi}|\alpha)$. Because we chose conjugate priors in Equation 3.21, we can marginalize out $\vec{\pi}$ (for details see [Griffiths and Ghahramani, 2005]):

$$P(C|\alpha) = \int \prod_{i=1}^{N} P(c_i|\vec{\pi})P(\vec{\pi}|\alpha d\vec{\pi}) \tag{3.25}$$

$$= \frac{\prod_{i=1}^{K}\Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \tag{3.26}$$

Taking the limit as $K \to \infty$,

$$P(C|\alpha) = \alpha^{K_+} \left(\prod_{k=1}^{K_+} (m_k - 1)! \right) \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \tag{3.27}$$

where $K_+, K_+ \leq N$ is the number of classes with at least one element, and m_k is the number of items in class k, and $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) = (\alpha - 1)!$ is the Gamma function.

We can now sample new values for Θ according to

$$P(\theta_k|C, X, \Theta_{-k}, \vec{\pi}, \alpha) \propto \prod_{i \text{ s.t. } c_i = k} P(x_i|c_i, \theta_k) P_{\mu_k, \Sigma_k}(\theta_k) \tag{3.28}$$

where $\Theta_{-k} = \{\theta_1, ..., \theta_{k-1}, \theta_{k+1}, ..., \theta_N\}$ and $P_{\mu_k, \Sigma_k}(\theta_k)$ is the probability of θ_k under μ_k, Σ_k.

Finally, C is sampled according to

$$P(c_i = k|C_{\neg i}, X, \Theta, \vec{\pi}, \alpha) \propto P(x_i|c_i, \Theta) P(c_i|C_{\neg i}) \tag{3.29}$$

where $C_{\neg i} = \{c_1, ..., c_{i-1}, c_{i+1}, ..., c_N\}$. We can then say

$$P(c_i = k|C_{\neg i}) = \begin{cases} \frac{m_k}{i-1+\alpha} & \text{if } k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & \text{if } k > K_+ \end{cases} \tag{3.30}$$

Note that this gives us a sampling method that handles an arbitrarily large number of classes, K, as, with some probability proportional to our Dirichlet process concentration parameter α, we add a new class $k > K_+$.

The Chinese Restaurant Process

As an illustration of how a Dirichlet process works, we consider a scenario known as the Chinese Restaurant Process (CRP). In the CRP scenario, patrons enter a Chinese Restaurant one by one. Inside the restaurant are infinitely many tables, T, and a customer chooses a table with probability

$$x_n|x_1, ..., x_{n-1} = \begin{cases} t_k & \text{with probability } \frac{num_{t_k}}{n-1+\alpha} \\ \text{new table} & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases} \tag{3.31}$$

where num_{t_k} is the number of patrons sitting at table t_k, and $k \in K_+$.

As you can see, this exhibits two main behaviors. First, the number of tables can grow infinitely, with probability proportional to α. Second, there is a clustering, or a "rich get richer" effect, in that new patrons are more likely to sit at tables that already contain large numbers of patrons. This is important, because it means that, as the data converges, the final distribution changes little [Aldous, 1985, Sudderth et al., 2005]. See Figure 3.2 for an illustration of the process.

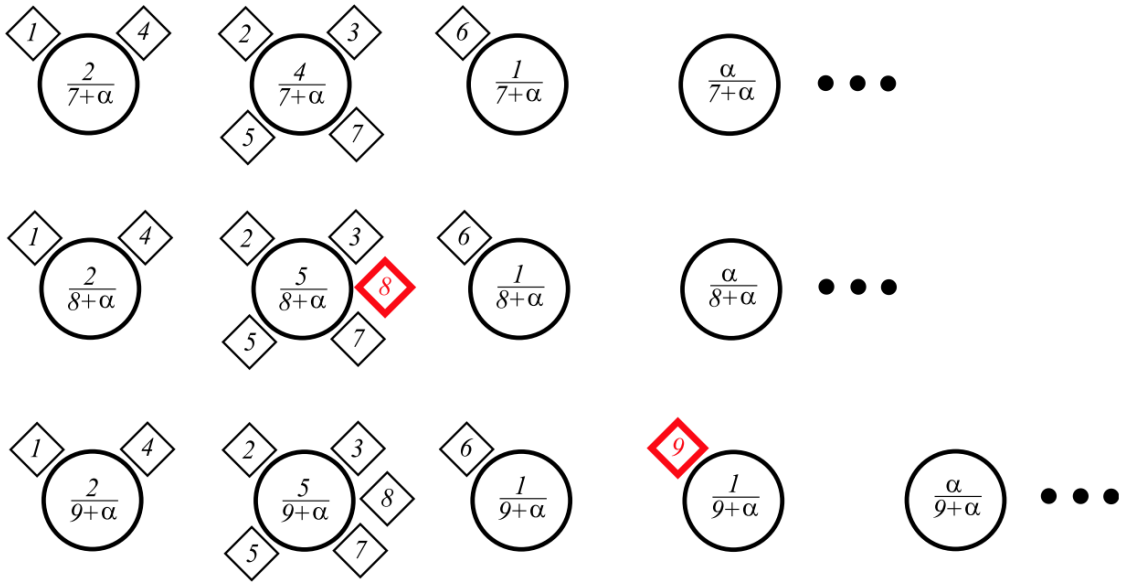


Figure 3.2: An example of the Chinese Restaurant Process, showing three steps of the process on each row. The probability of a new point choosing each table is shown in the center of the table. From [Sudderth et al., 2005]

Inference in Dirichlet Processes

Inference in Dirichlet process models has often been done using Markov Chain Monte Carlo (MCMC) techniques. An example of this is the Metropolis-Hastings approach for probability density estimation. In this approach a surrogate function, $f(x)$, is chosen, such that is proportional to the true density and easy to sample. Samples are then selected in an iterative fashion in which the value of the next point, x_{i+1}, is conditioned on the value of the current point, x_i. This is done by drawing the new point from a symmetric function, such as a Gaussian with mean x_i. An acceptance ratio is then calculated in which the function $f(x)$ is evaluated at x_i and x_{i+1}:

$$a = \frac{f(x_{i+1})}{f(x_i)} \tag{3.32}$$

Because $f(x)$ is proportional to the true underlying probability distribution, if this ratio is greater than 1 then x_{i+1} is thought to be more likely than x_i and the method is repeated centered around x_{i+1}. If a is less than 1, point x_{i+1} is chosen with probability a and x_i is chosen with probability $1-a$. An extension to this is Metropolis-within-Gibbs sampling, which deals with MCMC estimation on multivariate functions. In Metropolis-within-Gibbs sampling only one dimension is evaluated at a time, in a similar manner to Metropolis-Hastings, and after each evaluation the dimension to be evaluated is alternated. We refer the reader to Neal et al. for a more comprehensive overview of MCMC methods applied to Dirichlet processes [Neal, 2007].

There have been a number of other developments in inference in latent class models, particularly in the field of variational inference [Blei and Jordan, 2006]. Unlike MCMC models which randomly sample the function space, variational inference can be deterministically calculated and requires no sampling. First, the true probability distribution, $p(y|x)$ in which x is the input data and y are the outputs, is approximated by a simpler distribution $q(x)$. Then we can lower-bound the log likelihood of the data and use gradient ascent or other methods to increase the lower bound to some deterministically-calculable maximum.

3.1.2 Adaptively Adding Classes Using Dirichlet Process Priors

In this section we discuss an application of Dirichlet processes to our AGMM model. Instead of using AIC or BIC to estimate the number of classes at each iteration, we use a Dirichlet process prior to estimate whether or not a newly-collected sample point should be added to an existing class or added to an entirely new class.

Our formulation assumes a similar structure to that of the Chinese Restaurant Process described in Section 3.1.1. In that scenario, the probability of patron x_n choosing a table is proportional to

$$x_n|x_1,...,x_{n-1} = \begin{cases} t_k & \text{with probability } \frac{num_{t_k}}{n-1+\alpha} \\ \text{new table} & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases} \tag{3.33}$$

In the CRP case, all tables and all patrons have equal probability, and thus the probability of being at a new table or an old table are only dependent on the number of patrons already at the restaurant and sitting at the table, respectively.

In our work we can leverage the mixture model probabilities in order to obtain a more accurate distribution of our data. To demonstrate this, we introduce the *Food Court Process*. Imagine that instead of a restaurant with up to an infinite number of tables, there instead is a food court with up to an infinite number of restaurants. Patrons enter the food court, then choose a restaurant at which to eat.

In the CRP example, patrons are equally probable and their choice of table is based entirely on the Dirichlet concentration parameter, α, and the number of patrons at each table. In the food court case, imagine instead the more realistic scenario where each patron has a preference of food, and individual groups of people are more or less likely to appear. In this case, the model can be updated to look like this:

$$x_n|x_1,...,x_{n-1} = \begin{cases} r_k & \text{with probability } \frac{num_{r_k}}{n-1+\alpha} P_{r_k}(patron|r_k) \\ \text{new restaurant} & \text{with probability } \frac{\alpha}{n-1+\alpha} P(patron) \end{cases} \tag{3.34}$$

in which the probability of a patron choosing an existing restaurant is proportional to the popularity of the restaurant, weighted by their preference to that type of food. If the patron does not like any of the restaurants that already have patrons, then it is more likely they will choose to eat at a new one. As an example, see Figure 3.3.

(a) The Cookie Monster enters

(b) The Cookie Monster chooses to eat cookies

(c) Homer Simpson enters

(d) Homer chooses a new restaurant

Figure 3.3: An example of the Food Court Process. a) The Cookie Monster enters and sees that this food court has cookies. b) He chooses the popular cookie store and eats there. c) Homer Simpson enters. d) He likes both cookies and pizza, but decides to eat at a donut shop instead.

This is exactly the process at work in our approach. As we sample points, our mixture model builds up a number of distributions based on the data in the sample points belonging to those mixtures. When a new sample point arrives, the algorithm must decide if it should add a point to an existing mixture or create a new one, using the aforementioned equation:

$$P(c \in \{C, new\}|x) \propto \begin{cases} \frac{m_c}{n-1+\alpha} P_c(x|c) & \text{if } c \in C \\ \frac{\alpha}{n-1+\alpha} P(x) & \text{if } c = \text{new} \end{cases} \tag{3.35}$$

for sample x, the set of mixtures C, and m_c being the number of points in mixture c. $P(x)$ is the prior probability of the sample and $P_c(x|c)$ is the probability of our sample belonging to mixture c, given by our mixture model. We refer to this approach throughout this dissertation as **AGMM-AddClasses**.

Intuitively, one might imagine that, because we are sampling the most diverse set of points we can find, we will always be likely to create new mixtures, as sampled points are selected based on their "distance" from existing mixtures. In reality, however, due to the noise in the system this is not always the case. Locations that look promising given the contextual data may only appear promising because of a large amount of noise in that location's measurement. After fully observing the point it may be the case that that location is not novel enough to warrant the creation of a new mixture.

A number of important questions arise out of this formulation that are necessary for the approach to perform well, namely how to calculate the prior of our data, $P(x)$, and how to calculate a reasonable value for α. An accurate estimate of $P(x)$ helps disambiguate outliers from more realistic data points. Prior work has done this through methods like density estimation, such as Vatturi

and Wong's proposal to use hierarchical mean shift to estimate $P(x)$ [Vatturi and Wong, 2009]. However, in our approach, we note that rare classes or outliers are actually one of the objectives of our sampling strategy, so our prior should not penalize such points. This is even more important in the case of the microscopic scenario discussed in Chapter 6, in which data is obtained one at a time and a proper density estimate cannot be easily calculated. We will discuss our assumptions on $P(x)$ in later chapters as it applies to each different dataset.

Next we consider a method for selecting an appropriate value for the Dirichlet concentration parameter α. Referring back to Equation 3.35, an appropriate value of α determines the likelihood of a point being a member of a new, undiscovered class. A high value of α leads to a larger number of new mixtures being created, while a smaller version of α leads to data being grouped with existing mixtures. If we select a static value for α, eventually the number of member points in existing mixtures, num_{t_k}, outweighs α and the likelihood of a new point being added to an undiscovered mixture diminishes. Instead, we use the Gibbs sampling approach of Esocbar and West [Escobar and West, 1995] that estimates an appropriate value of α at each time step, given the number of points sampled and the number of classes. They propose placing a prior on α using a gamma distribution, $\Gamma(a, b)$. Their method then samples a quantity, η, given the current estimate of α, then resamples α given the current value of η. η, given a value of α, is drawn from a Beta distribution:

$$\eta|\alpha, k, n \sim \beta(\alpha + 1, n) \tag{3.36}$$

where k is the number of mixtures in our model so far, and n is the number of observed sample points. α given η is then a mixture of two gamma distributions:

$$\alpha|\eta, k, n \sim \pi\Gamma(a + k, b - \log(\eta)) + (1 - \pi)\Gamma(a + k - 1, b - \log(\eta)) \tag{3.37}$$

where the ratio of the mixture is determined by

$$\frac{\pi}{1 - \pi} = \frac{a + k - 1}{n(b - \log(\eta)} \tag{3.38}$$

We then use Gibbs sampling to estimate a value of α. After an initial burn-in of 128 sample steps [Geman and Geman, 1984], we iterate another 128 steps to find an estimate of α. We use a weak prior on α, $\Gamma(1, 1)$, and through multiple tests of different parameters for this gamma prior have found this to work reasonably well.

As an aside, Escobar and West offer an interesting and very thorough look at the relationship between α and k. They note that, for n moderately large (50 to 250 samples),

$$E(k|\alpha, n) \approx \alpha \ln\left(1 + \frac{n}{\alpha}\right) \tag{3.39}$$

A value of $\alpha = 1$ corresponds to a prior receiving the weight of one observation in the posterior, leading to an expected value of k in the single digits for n moderately large. We encourage the

reader to refer to Escobar and West [Escobar and West, 1995] for a more thorough look at the relationship between these parameters.

An important alternative approach is the Dirichlet Process Guassian Mixture Model (DPGMM) [Rasmussen, 1999]. This model evaluates the dataset to estimate the number of observed classes, then builds a Guassian mixture model using that estimate. We evaluated this model in earlier tests and found that it produced poor results when used in our adaptive sampling application. This is because it is very sensitive to the the value of α used in its construction when trained with a low number of samples. Different values of α produced wildly different results, and a single value of α did not perform consistently well from experiment to experiment.

However, the DPGMM approach might be a good way to refine the AGMM model. Over time, points collect within each AGMM mixture, but it may be beneficial to "reshuffle" these points periodically as new points are collected. In this case, a value of α can be estimated from the existing distribution, and the DPGMM approach can be used to potentially improve the AGMM mixtures.

3.1.3 Misclassification Estimation Using Dirichlet Priors

Haines et al. describe an alternate approach to sample selection, discussed in more detail in Section 4.5.1. In their work, they use a similar mixture model approach for density estimation and class membership, yet they are working in an active learning scenario in which observing a sample point provides the true underlying class label. They are thus less worried about creating mixtures, as the mixture memberships are provided upon sampling, but their overall methods of sample selection are relevant here.

Similar to what we described in Equation 5.5, they are interested in calculating the probability of any point being incorrectly classified given an existing model. This has been considered in prior work, such as Lewis and Gale's work on text classification [Lewis and Gale, 1994], in which they propose selecting sample points with class probabilities closest to 0.5. However, their work doesn't consider the probability of a point belonging to an unobserved case.

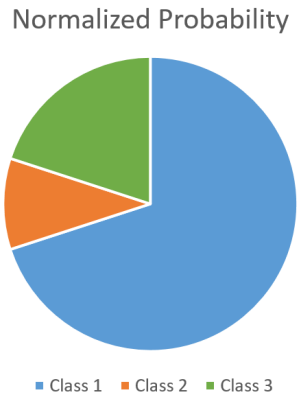
Haines et al. instead considers what they describe as $P(Wrong)$, or the probability of being incorrectly classified given all existing classes, as well as the underlying probability that it belongs to a new class:

$$P(Wrong|x) = 1 - P_n(c'|x) \tag{3.40}$$

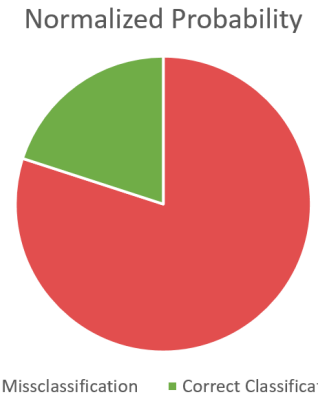
$$c' = \arg\max_{c \in C} P_c(x|c) \tag{3.41}$$

where P_n is the Dirichlet-weighted probability of belonging to an existing class or a new, unobserved class:

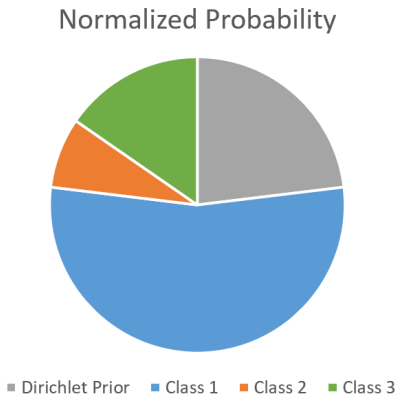
$$P_n(c \in \{C \cup new\}|x) \propto \begin{cases} \frac{m_c}{n-1+\alpha} P_c(x|c) & \text{if } c \in C \\ \frac{\alpha}{n-1+\alpha} P(x) & \text{if } c = \text{new} \end{cases} \tag{3.42}$$



(a) Normalized probabilities of a sample point coming from each mixture.



(b) Probability of misclassifying Class 3.



(c) Normalized probabilities, including a Dirichlet prior estimating the probability of a new class.

(d) Updated probability of misclassifying Class 3.

Figure 3.4: An example of the weighting process described in Equation 5.5. a) The probability of the sample point belonging to each class is calculated and normalized to sum to 1. b) The probability of incorrectly classifying a point that belongs to Class 3, shown in red. c) Normalized probabilities, including the probability of the sample belonging to a new class according to our Dirichlet prior, P_{Dir}. d) Updated probability of misclassification for a point belonging to Class 3.

Similar to what we described in Section 3.1.2, they use the technique proposed by Escobar and West [Escobar and West, 1995] for estimating the Dirichlet concentration parameter α. They also assume a uniform prior for $P(x)$.

Finally, they note that it is not always advantageous to select the point with the highest probability of being wrong. Instead, they see improved results with a soft selection approach, in which

a handful of the points most likely to be misclassified are probabilistically selected with weights proportional to their probability of misclassification.

We apply a modification of this technique to our sample selection process, referring to it as **AGMM-P(Wrong)**. We consider the case where our mixture model is defined as the AGMM model discussed above. When points are considered for sampling, we use our existing mixtures to calculate the probability of misclassification given existing mixtures, weighted by the Dirichlet probability of the point belonging to a new class. Similar to Haines, we select the point that maximizes $P(Wrong)$ as described in Equation 3.40. The selected point is then added to our AGMM model, as described in Section 3.1, using a fixed number of classes in the AGMM model.

When compared to Haines et al., we make a number of modifications to improve his results for our scenario. First, Haines discusses using a soft selection strategy, in which the point most likely to be wrong is not selected, but instead a subset of points with high misclassification probabilities is considered, and one is chosen randomly based on its misclassification probability when weighted against all others in the subset. In our work, we find that using a soft selection strategy like this under-performs compared to selecting the point with a maximum value of $P(Wrong)$, so our results are calculated by selecting the point that is most likely to be misclassified. Second, Haines considers classification, in which observing a point gives a true class label. In our work we see less noisy versions of the data, and not a true class label. This is handled somewhat by our AGMM model, but we find that our scenario is inherently much more difficult, as ambiguities in the model creation process tend to have a large impact in the overall selection process, especially in the calculation of $P_c(x|c)$. To improve this, we had to normalize the data to be zero mean and unit variance, for similar reasons to what we described in Section 3.1.2. This leads to a better estimate of $P_c(x|c)$, especially when dealing with a large number of dimensions in our data. Similar to Haines, we use a uniform prior for $P(x)$, the equation for which is shown in Equation 5.7. These changes make it possible for our Dirichlet prior and class probabilities to be in the same scale, enabling an accurate calculation of $P(Wrong)$.

3.2 One-Step Bayesian Optimization

In addition to our mixture model work, we introduce an alternative sample selection process that draws from Bayesian optimization approaches to find new sample selection points.

Bayesian optimization is a field of research focused on using Bayesian methods to model reward functions, then sampling methodically within those models to maximize reward. Typically this is formulated as an optimization problem of some nonlinear function $f(x)$ over a compact set A [Brochu et al., 2010b]:

$$\max_{x \in A} f(x) \tag{3.43}$$

In most optimization cases, this objective function is typically assumed to be convex, known, or easy to sample, such that repeated queries are inexpensive. Realistically, this is not often the case, and methods like Bayesian optimization have been developed to allow for indirect optimization.

Bayesian optimization is applicable when the direct function to optimize is difficult to model or unknown, or evaluations are expensive, but when samples can be collected at given values in order to estimate the structure of the underlying function. It operates on Bayes' Theorem, using the samples collected as evidence to update the posterior estimate of an approximation to the underlying true function. It has been shown to be very efficient in terms of the number of samples required to sample a function [Mockus, 1994, Jones et al., 1998, Streltsov and Vakili, 1999, Jones, 2001, Sasena, 2002].

In more detail, Bayesian optimization is interested in finding the posterior $P(f|D_{1:t})$, given prior samples $D_{1:t} = \{x_{1:t}; f(x_{1:t})\}$, then sampling from that posterior methodically in order to estimate the next location, $x_{t+1} \in \mathcal{A}$, the next-best estimate of the maximum of the underlying objective function. Typically the unknown function is modeled by a known *surrogate function* or *response surface*. This is typically done by using the posterior mean of a Gaussian process model [Rasmussen, 2006]. Gaussian process models can intuitively be thought of as a function that returns a Gaussian with a mean and variance at every point. As data is collected the Gaussian process is recalculated and the mean and covariance functions adjust to fit to the new data, typically providing greater accuracy in regions that have been sampled. An example of the process is shown in Figure 3.5.

Sampling within Bayesian optimization is a trade-off between exploration, in this case sampling unknown regions of the response surface, and exploitation, sampling regions of the response surface known to be promising. The choice of where to sample is typically handled by an *acquisition function*, or some estimate of the "reward" of sampling at a given location. It is thus the goal of many acquisition functions to best take advantage of the mean and variance estimates from our response surface in order to choose sampling points that both reduce the uncertainty in the model and have a high likelihood of improving upon the estimate of the objective function's maximum. Optimizing this function typically is done using the principles of maximum expected utility, or minimum expected risk [Brochu et al., 2010b]. Examples of these acquisition functions are detailed below in Section 3.2.

It is important to note that Bayesian optimization typically operates on a single, static objective function. This function does not change over time or with observations, and, thus, repeated queries into the function build upon prior knowledge and can be used to improve the accuracy of the response surface. As we will detail in later sections of this thesis, this is often not the case in our scenario, as our objective function is the potential reward from sampling a site. Sampling at the chosen site affects that reward, necessitating a re-estimation of the response surface. Thus, our acquisition functions and the paramters chosen for our response surface need to keep this in mind and focus on reliably maximizing the reward of a single sample.

One-Step Bayesian Optimization

In our work we consider a Bayesian optimization approach to selecting sample points. We consider the problem of selecting the next sample point with the highest expected entropy, x_{n+1}, given our

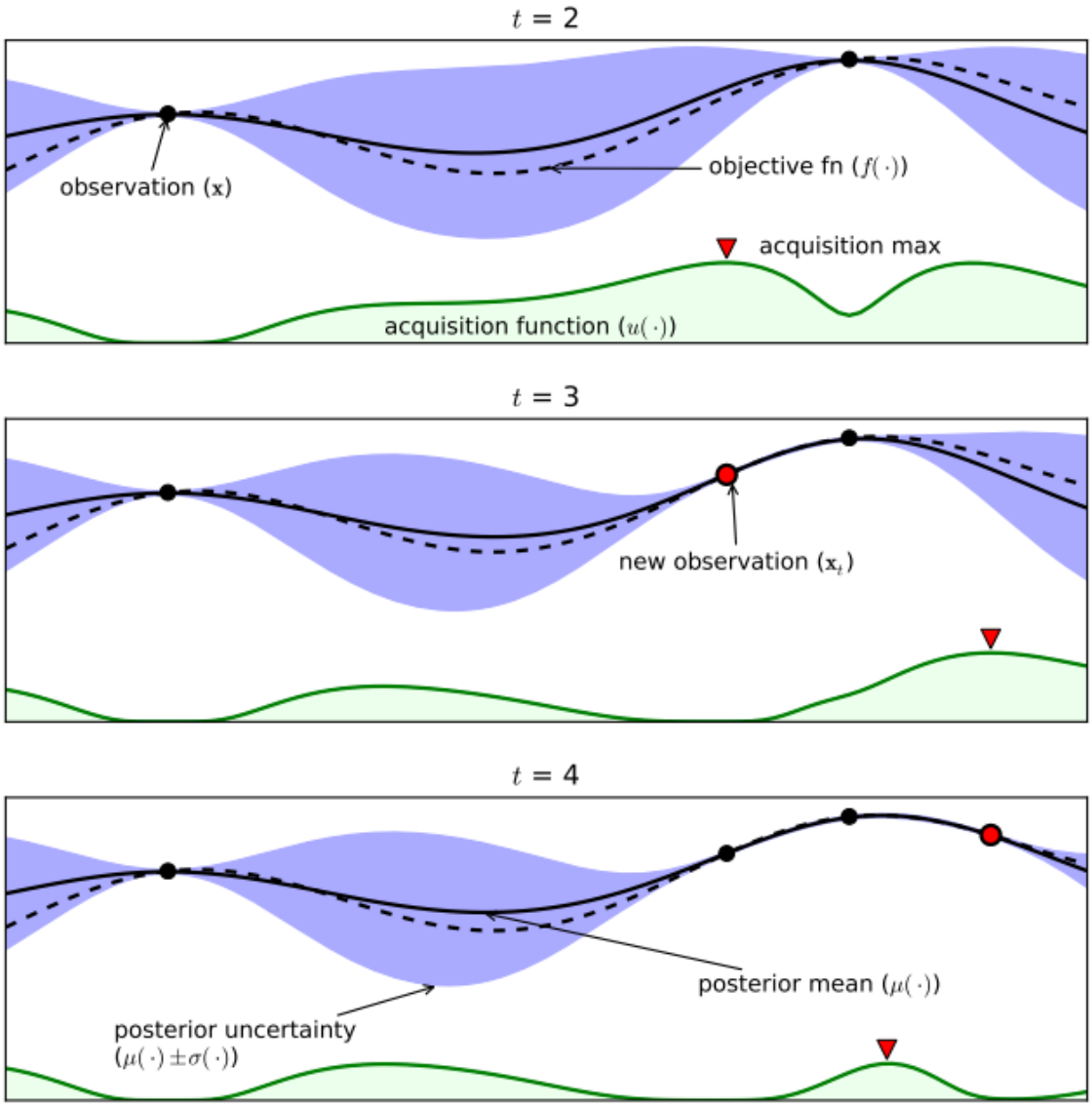


Figure 3.5: An example of the Bayesian optimization process using a 1-dimensional Gaussian process model. The objective function, shown in dashed black lines, is the underlying function that is difficult to estimate but can be sampled repeatedly. A Gaussian process is used to estimate the underlying function. Its mean function is shown as a solid black line and variance is shown as a shaded blue region. The acquisition function is shown in green. This is an estimate of the reward of sampling at any point along the x-axis, and typically the maximum value of this function is sampled at each iteration. From [Brochu et al., 2010b].

sample set X:

$$x_{n+1} = \arg\max_{x_i} H(X \cup x_i), \ \ \forall x_i \notin X \tag{3.44}$$

However, the calculation of entropy is intrinsically dependent on the set of samples:

$$H(X) = \frac{1}{2} \sum_{b \in B} \log(2\pi\sigma_b) \tag{3.45}$$

where $b \in B$ are each of the elemental channels. This means that our target objective function is changing at each sampling step, a major difference from typical Bayesian optimization approaches. Thus, our objective is to find acquisition functions that provide the greatest return over a single step, what we call *one-step Bayesian optimization*.

Because we are not building up and iterating upon a single response surface over time as is typically done, the Gaussian process model's benefits to our work are twofold. Most importantly, it provides a smoothing over the data, ameliorating the effects of noise, especially in dense regions of feature space. Additionally, the variance estimates provide a measure of uncertainty not available in many other methods that allow us to consider explicitly the trade off between true exploration or exploitation. Comparing to our AGMM approach, in which we have an estimate of the distance from existing mixtures, but here we also have a measure of the uncertainty in our estimate.

Selection of Gaussian Process Parameters

In order to provide an accurate estimate of our objective function, we require two important steps. First, training of Gaussian processes are $\mathcal{O}(n^3)$ in run time and $\mathcal{O}(n^2)$ space in memory. As our sample space is inherently large, approximately 30,000 potential sample points, we must choose a subset of training data for our Gaussian process model. This is an important step, as the success of the acquisition functions depends on how accurately the Gaussian process estimates the scene. A poor selection of training points, eg. a training set made up of the least-diverse points, will lead to a sampling strategy that is dominated by pure extrapolation and will likely lead to poor performance. If the training points are good, and are representative of the feature space, especially those sections of the feature space that have high reward, performance is much better and the acquisition functions are much more accurate.

There are a handful of approaches that have dealt with the sparsification of training sets in scenarios with high volumes of training data. These sparsification methods typically reduce the runtime to $\mathcal{O}(nm^2)$ and memory required to $\mathcal{O}(nm)$, where m is a user-defined limit on the amount of training data. As Candela and Rasmussen discuss [Candela and Rasmussen, 2005], the selection of these m points then becomes an exercise in how to select representative points from the prior distribution.

The two most popular sparsification methods are the Fully Independent Training Conditional (FITC) [Snelson and Ghahramani, 2005] and Partially Independent Training Conditional (PITC)

[Candela and Rasmussen, 2005] algorithms. Both of these approaches select a subset of "induction" points, or points taken from within the training data, or some approximation of the data, that are aimed to provide a representative coverage of the data itself. For a very thorough look at both methods, we refer the reader to [Snelson, 2007].

FITC works by selecting a set of pseudo-inputs that span the training space. This is done by maximizing the marginal likelihood of the pseudo-inputs and the hyperparameters of the Gaussian process using gradient ascent [Snelson, 2007].

Alternatively, PITC assumes local "blocks" of data. Where FITC assumes complete conditional independence between training points, PITC assumes conditional independence only between blocks. It is then a matter of determining which points to use in the diagonal blocks. As the blocks should be of similar points, clustering methods are commonly used. In farthest point clustering [Gonzales, 1985], an initial seed point is selected, then additional clusters are created using the point farthest from the existing clusters. Alternatively, clusters can be chosen randomly. At test time, points are assigned the covariance of their closest cluster.

While we have tested these methods in our approach, we have found that random sampling has actually outperformed these more computationally-intensive methods. In our tests we therefore select a random subset of 400 points at each training iteration. This is likely because 400 points typically spans the feature space, providing an adequate representation of the full data. Additionally, we tried clustering methods, such as mean shift [Keinosuke and Hostetler, 1975] or k-means [Lloyd, 1982], but found that a random selection of points performs as well or better, with a much lower computational cost.

The second important step in creating an accurate Gaussian process is in the selection of the mean function and a covariance kernel. The general formulation of a Gaussian process is:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{3.46}$$

where $m(x)$ is the mean function and $k(x, x')$ is the covariance kernel. Mean functions encode assumptions about the overall structure of the data, and, without loss of generality, can be considered zero. Alternative priors on the mean can be found in [MartinezCantin et al., 2009] and [Brochu et al., 2010a].

Covariance kernels are typically more important to the overall performance of the Gaussian process, and are functions that characterize the similarity of two arbitrary points, x and x', with some restrictions on their structure. A *stationary* covariance kernel is a function on $x - x'$, and are thus invariant to translations in input space. *Isotropic* covariance kernels are defined on $|x-x'|$, and are thus invariant to all rigid motions. Stationary covariance kernels are appropriate when it can be assumed that there is relative smoothness within the feature space. Nonstationary covariance kernels are more appropriate if there are discontinuities in the data or sharp boundaries.

An example of a covariance kernel is the commonly-used squared exponential kernel:

$$k_{SE}(r) = \exp\left(-\frac{r^2}{2\ell}\right) \tag{3.47}$$

where $r = |x - x'|$ and ℓ is the characteristic length scale, a smoothing parameter.

The Matern class of covariance functions are stationary covariance functions with a smoothing parameter $\nu > 0$ and a length-scale parameter $\ell > 0$. In many cases, including ours, ℓ is chosen to be a scalar, making the covariance function isotropic as well.

$$k_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{\ell} \right) \tag{3.48}$$

where K_ν is a modified Bessel function [Abramowitz and Stegun, 1965]. If ν is of the form $\nu = p + \frac{1}{2}$, where p is a non-negative integer, the function is greatly simplified. In many machine learning applications, ν typically chosen to be one of $\frac{1}{2}, \frac{3}{2}$, or $\frac{5}{2}$. For $\nu = \frac{1}{2}$, the function is fairly rough, and many optimizers likely have difficulty discerning values of $\nu \geq \frac{7}{2}$ from noisy data [Rasmussen, 2006]. For $\nu = \frac{3}{2}$, and $\nu = \frac{5}{2}$, the functions simplify to

$$k_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3} r}{\ell} \right) \exp \left(- \frac{\sqrt{3} r}{\ell} \right) \tag{3.49}$$

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5} r}{\ell} + \frac{5 r^2}{3 \ell^2} \right) \exp \left(- \frac{\sqrt{5} r}{\ell} \right) \tag{3.50}$$

Stein argues that the Matern kernel is more appropriate for physical data than the squared exponential kernel [Stein, 1999]. He believes that the smoothness assumptions made by the squared exponential kernel are unrealistic for physical data, and that the irregularities inherent in physical data is more easily handled using the Matern kernel.

In many cases, ours included, the feature space is sparse. Thus, it may be advantageous to have covariance kernels with compact support, or the property that they evaluate to zero when the distance between two points is greater than some threshold, leading to computational advantages. One kernel of this type is the piecewise polynomial kernel, examples of which are:

$$k_{pp,q=1}(r) = \sigma^2 \max(1 - r, 0)^{j+1}((j+1)r + 1) \tag{3.51}$$

$$k_{pp,q=2}(r) = \sigma^2 \max(1 - r, 0)^{j+2}((j^2 + 4j + 3)r^2 + (3j + 6)r + 3)/3 \tag{3.52}$$

$$j = \left\lfloor \frac{D}{2} \right\rfloor + q + 1 \tag{3.53}$$

where D is the dimension of the data. As covariance kernels must be positive definite, the structure of these functions and their derivation are complicated, yet they are efficient to compute. More discussion on their derivation can be found in [Wendland, 2009].

The hyperparameters of the covariance kernel are typically learned by maximizing the log likelihood of the training data [Jones et al., 1998, Sasena, 2002, Rasmussen, 2006], improved with the aid of informative hyperpriors on the hyperparameters [Lizotte, 2008]. For more discussion on this topic we refer the reader to Rasmussen's excellent coverage of the topic [Rasmussen, 2006].

In this work we focus on stationary kernels, particularly the Matern and piecewise polynomial kernels listed above. We limit ourselves to stationary kernels, as we assume that there is smoothness within our feature space. Furthermore, initial experimentation indicated that these two kernels were the most promising on our dataset.

Acquisition Functions

Given an accurate response surface, the Bayesian optimization problem then becomes a selection of an acquisition function. Acquisition functions are functions that guide the search for the optimum in the objective function. Typically, these functions observe the mean and variance of the response surface at candidate measurement locations, then compute a score for these locations based on their expected value, uncertainty, or a mix of both. A high value in the acquisition function corresponds to the prediction of a high value in the objective function, so we choose sample points that maximize our acquisition functions at each step:

$$x_{n+1} = \arg\max_{x_i} u(x_i|X) \tag{3.54}$$

where $u(\cdot)$ is the generic representation for an acquisition function and X are the samples collected thus far.

In this work we consider each of acquisition functions discussed in this section. In many of these functions, we consider the improvement over an *incumbent*, $f(x^+)$, the highest value observed in our objective function thus far. Typically, when multiple samples are collected from the same objective function, this is considered $x^+ = \arg\max_{x_i \in X} f(x_i)$. In our case, where the objective function changes at each iteration, we instead consider $x^+ = \arg\max_{x_i \in X_{unsampled}} \mu(x_i)$, the unsampled point with the highest predicted value under our current model.

The first acquisition function we consider is that of Probability of Improvement (PI) [Kushner and Yin, 1997]. PI maximizes the probability of improving over the incumbent value:

$$\begin{aligned} PI(x) &= P(f(x) \geq f(x^+)) && (3.55)\\ &= \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right) && (3.56)\end{aligned}$$

where Φ is the cumulative distribution function.

The problem with this formulation is that points with high probability of being infinitesimally greater than x^+ will always be chosen ahead of points with a slightly lower score but higher uncertainty. Instead, a modification of the algorithm adds a trade-off parameter, ξ:

$$\begin{aligned} PI(x) &= P(f(x) \geq f(x^+) + \xi) && (3.57)\\ &= \Phi\left(\frac{\mu(x) - f(x^+) - \xi}{\sigma(x)}\right) && (3.58)\end{aligned}$$

This trade-off parameter makes PI slightly less greedy, in the fact that it selects points most likely to offer an improvement of at least ξ. In our tests we use a value of $\xi = 0.01$, but others propose a schedule for ξ that decays over time [Kushner and Yin, 1997]. A handful of papers have studied this parameter more extensively [Torn and Zilinskas, 1989, Jones, 2001, Lizotte, 2008]. We refer to this approach as **GP-PI**.

While this method is straightforward, it often performs very well. Jones [Jones, 2001] notes that the performance of PI

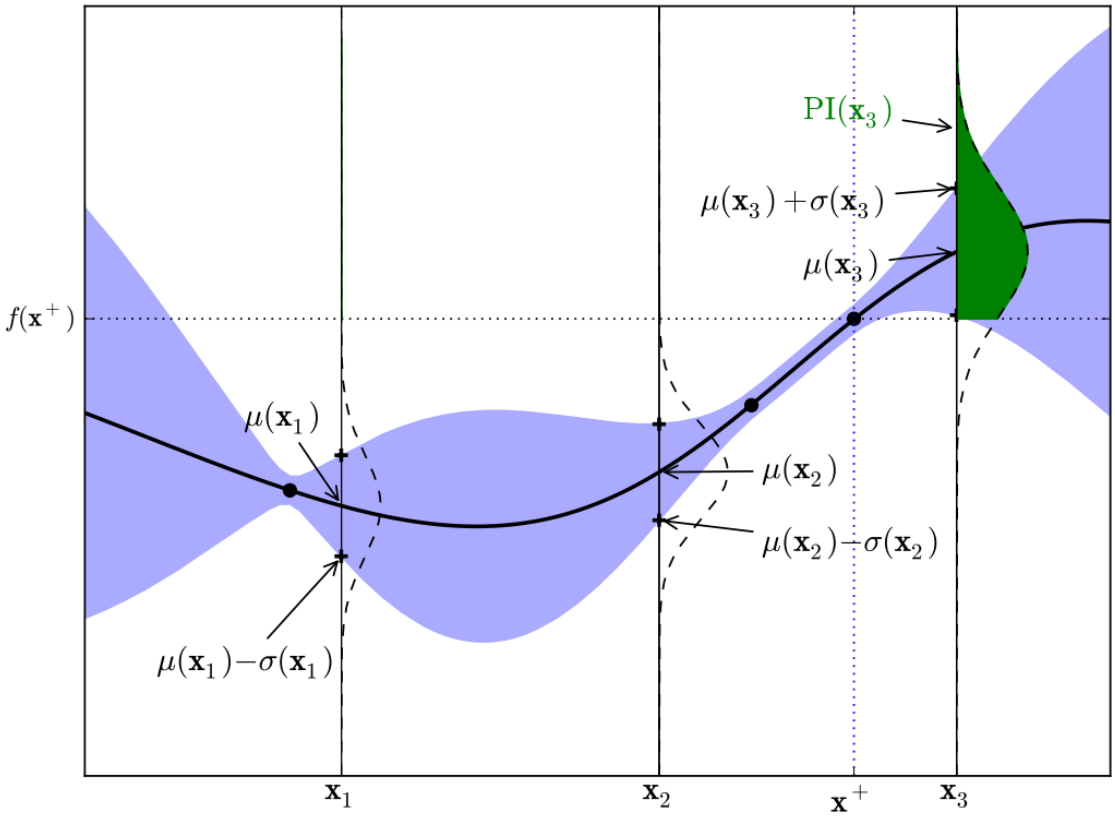


Figure 3.6: An example of the probability of improvement (PI) algorithm. The dark line and shaded area represent the mean and variance of the posterior of the Gaussian process. The green shaded area represents a measure of improvement over the incumbent, $f(x^+)$. From [Brochu et al., 2010b].

> "is truly impressive. It would be quite natural if the reader, like so many others, became enthusiastic about this approach. But if there is a single lesson to be taken away from this paper, it is that nothing in this response-surface area is so simple. There always seems to be a counterexample. In this case, the difficulty is that [the PI method] is extremely sensitive to the choice of the target. If the desired improvement is too small, the search will be highly local and will only move on to search globally after searching nearly exhaustively around the current best point. On the other hand, if [ξ] is set too high, the search will be excessively global, and the algorithm will be slow to fine-tune any promising solutions."

As an alternative, consider not keeping ξ fixed and instead calculating the probability of im-

proving over $f(x^+)$ by an amount I. Here the likelihood of improvement is calculated as:

$$\frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(-\frac{(\mu(x) - f(x^+) - I)^2}{2\sigma^2(x)}\right) \tag{3.59}$$

and then the expectation of this improvement (EI) as:

$$E(I) = \int_{-\infty}^{\infty} I \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(-\frac{(\mu(x) - f(x^+) - I)^2}{2\sigma^2(x)}\right) dI \tag{3.60}$$

This can be computed analytically as:

$$EI(x) = \begin{cases} (\mu(x) - f(x^+))\Phi(Z) + \sigma(x)\phi(Z) & \text{if } \sigma > 0 \\ 0 & \text{if } \sigma = 0 \end{cases} \tag{3.61}$$

$$Z = \begin{cases} \frac{\mu(x) - f(x^+)}{\sigma(x)} & \text{if } \sigma > 0 \\ 0 & \text{if } \sigma = 0 \end{cases} \tag{3.62}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ indicate the CDF and PDF of the standard normal distribution, respectively [Mockus et al., 1978, Jones et al., 1998].

While PI offers a more greedy approach to sampling, EI offers a more balanced approach between exploration and exploitation. If the variance of a candidate point is large, we explore, but if the mean of a candidate point is large, we exploit. If we instead want to encode a trade-off between these behaviors, we can again include a trade-off parameter, ξ, suggested by Lizotte [Lizotte, 2008]:

$$EI(x) = \begin{cases} (\mu(x) - f(x^+) - \xi)\Phi(Z) + \sigma(x)\phi(Z) & \text{if } \sigma > 0 \\ 0 & \text{if } \sigma = 0 \end{cases} \tag{3.63}$$

$$Z = \begin{cases} \frac{\mu(x) - f(x^+) - \xi}{\sigma(x)} & \text{if } \sigma > 0 \\ 0 & \text{if } \sigma = 0 \end{cases} \tag{3.64}$$

Lizotte finds good performance with a value of $\xi = 0.01$ and finds that a decaying schedule, as suggested in the PI literature, actually under-performs when compared against using a static value for ξ. We refer to this approach as **GP-EI**.

As an alternative to improvement-based acquisition functions, we also explore the use of functions on confidence bounds. These were first proposed by Cox and John [Cox and John, 1992, Cox and John, 1997], with the name Sequential Design for Optimization (SDO). SDO chooses points using the lower confidence bound:

$$LCB(x) = \mu(x) - \kappa\sigma(x) \tag{3.65}$$

where $\kappa \geq 0$. They were interested in minimization, finding the points with a high expected value, penalized by uncertainty, but the function is typically reversed and instead maximized using an upper confidence bound (UCB) metric:

$$UCB(x) = \mu(x) + \kappa\sigma(x) \tag{3.66}$$

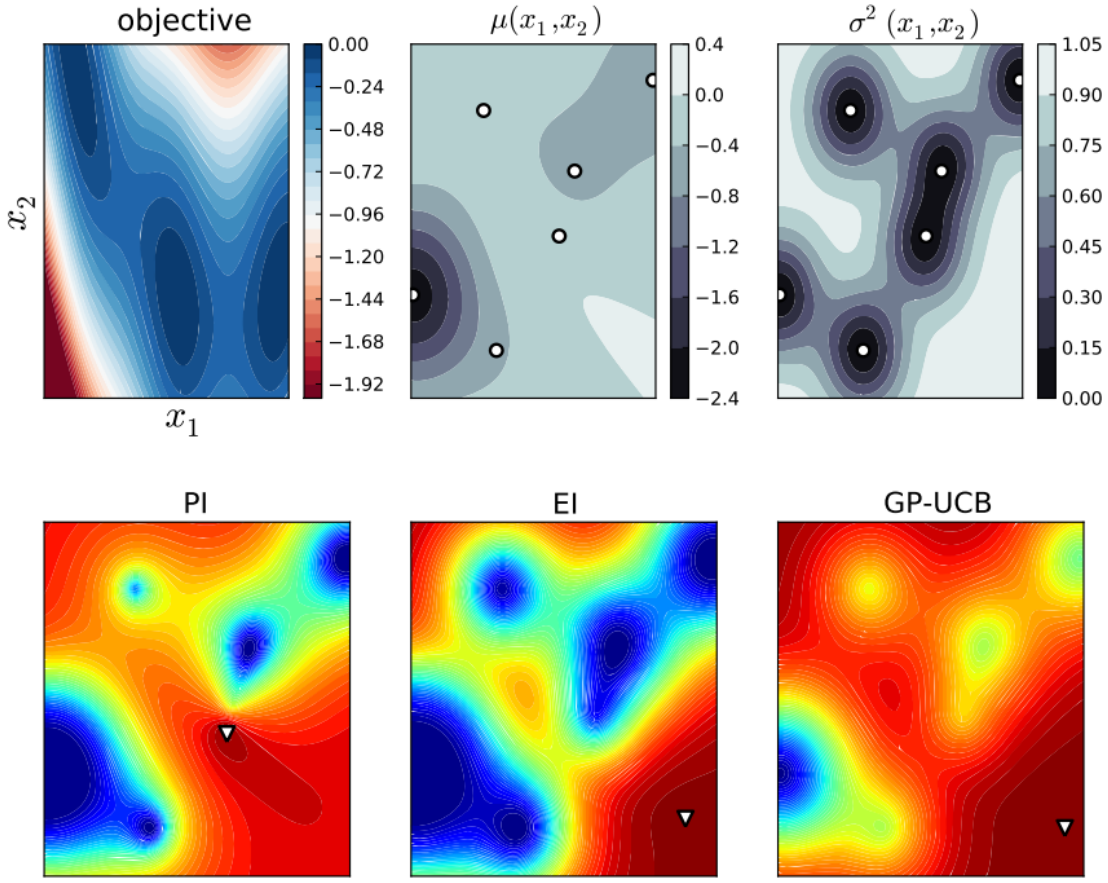


Figure 3.7: A demonstration of acquisition functions and their settings in 2 dimensions. The top left image shows the objective function. The top center and top right images show the mean and variance of the posterior Gaussian process model, respectively. The bottom images show the acquisition functions and their maximums for probability of improvement (bottom left), expectation of improvement (bottom center), and upper confidence bound (bottom right).

Maximizing this finds points that have a high mean and a potentially high reward, depending on the uncertainty.

Srinivas et al. propose a slight change to this formulation, formulating it as a multi-armed bandit problem instead of Bayesian optimization [Srinivas et al., 2010]. In that case, one can define the acquisition function as the instantaneous regret function:

$$r(x) = f(x^*) - f(x) \tag{3.67}$$

where $f(x^*)$ is the reward of the best possible point, and the regret measures how close the selected point x, was to the optimal point.

The goal is them to minimize the regret over T timesteps:

$$\min \sum_{t}^{T} r(x_t) = \max \sum_{t}^{T} f(x_t) \tag{3.68}$$

Returning to the UCB function defined in Equation 3.66, Srinivas et al. set $\kappa_t = \sqrt{\nu\tau_t},\ \nu > 0$ and maximize

$$GP - UCB(x) = \mu(x) + \sqrt{\nu\tau_t}\sigma(x) \tag{3.69}$$

With $\nu = 1$ and $\tau_t = 2\log(t^{\frac{d}{2}+2}\frac{\pi^2}{3\delta})$, for $\delta \in (0, 1)$, Srinivas et al. show that GP-UCB is *no regret* with high probability, meaning that the regret over time goes to zero:

$$\lim_{T\to\infty} \frac{R_T}{T} = 0 \tag{3.70}$$

$$R_T = \sum_{t=1}^{T} f(x^*) - f(x) \tag{3.71}$$

This implies a lower bound on the convergence rate of the optimization. We refer to this approach as **GP-UCB**.

Figures 3.7 and 3.7 show examples of the PI, EI, and GP-UCB acquisition functions in two scenarios.

Finally, we explore two simple models. The first, **GP-MaxMean**, samples from the point with the maximum expected value in our Gaussian process:

$$x_{n+1} = \underset{x_i \in X_{unsampled}}{\arg\max}\ \mu(x_i) \tag{3.72}$$

This is largely a measure of the ability of the Gaussian process to estimate the reward. Additionally, it acts as a smoothing function to reduce the impact of noise in sample-dense regions of our feature space.

The second model, **GP-MaxVar**, is a simple variance-based acquisition function aimed to specifically achieve the goals of our scenario. Here we simply select the point with the highest variance in our model [Lewis and Gale, 1994]:

$$x_{n+1} = \underset{x_i \in X_{unsampled}}{\arg\max}\ \sigma(x) \tag{3.73}$$

where $\sigma(x)$ is the variance estimate calculated using our Gaussian process. This acquisition function is designed to find points with the greatest uncertainty in our model. Because Gaussian processes have the property that the posterior variance is independent of observations collected thus far, one can measurement locations with the greatest uncertainty and even choose the subset of points to be sampled in advance, due to this independence.

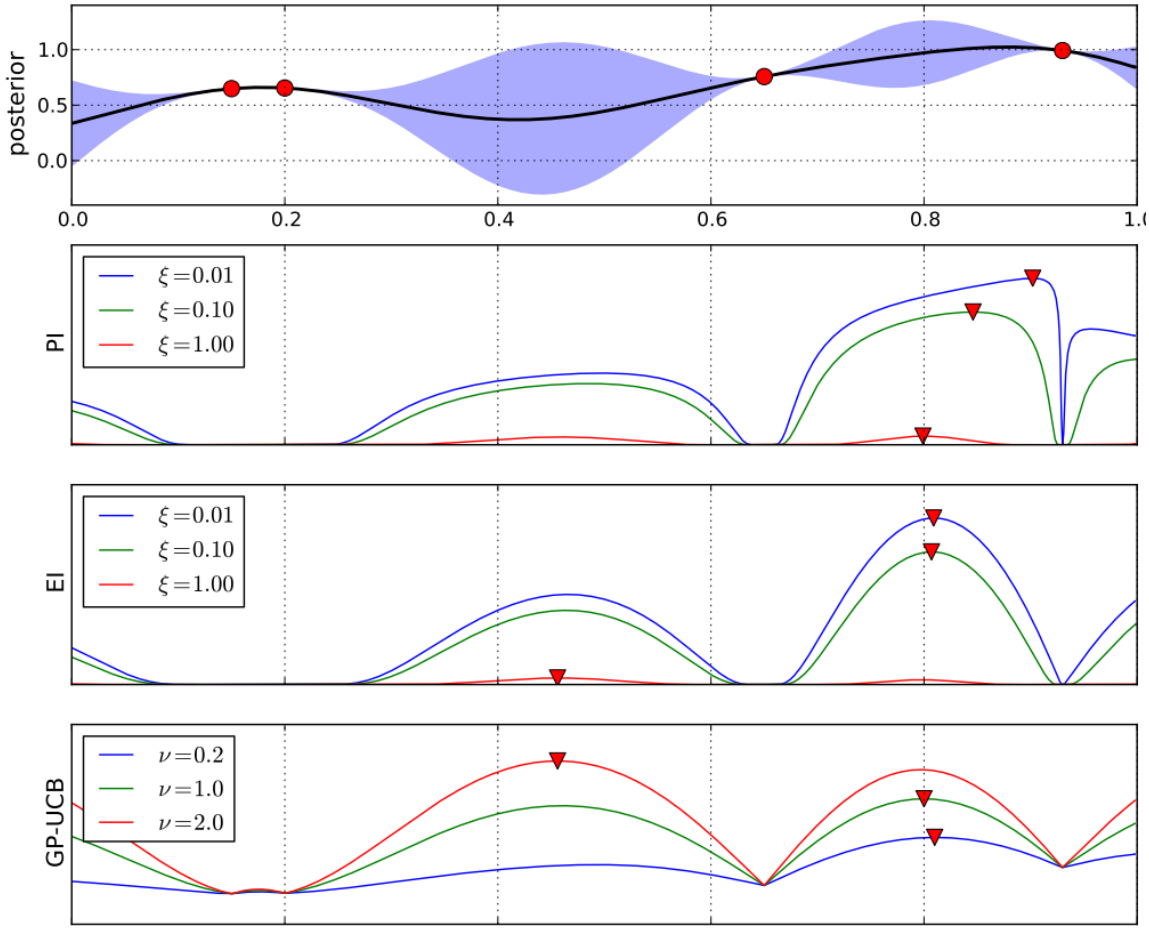


Figure 3.8: An example of three different acquisition functions, PI, EI, and GP-UCB. Triangles indicate the maximum values for each of the functions and their specified parameters. The top graph shows the posterior of the Gaussian process, in which the dark line shows the mean and the shaded area represents the variance at that x location. From [Brochu et al., 2010b].

In both of these simplified models, if our Gaussian process is trained with a reasonable selection from our prior distribution, then the high mean and high variance points are likely to offer a large amount of new information to our sample set. However, this method does closely tie to how representative our training points are of the true prior. A poor selection of training points leads to inaccurate uncertainty estimates and a poor acquisition function estimate.

3.3 Scoring Metrics

We score our results using two main metrics: entropy and reconstruction error. These metrics are chosen to be complementary. Entropy quantifies the amount of diversity in our samples, while reconstruction error quantifies the "coverage" of our sample set.

A high entropy score means that there is a large amount of information in the samples. We calculate this using differential entropy:

$$H(x) = \frac{1}{2}\sum_{b\in B}\log(2\pi\sigma_b) \tag{3.74}$$

where $b \in B$ are each of the channels in our data. This score rewards samples that increase the overall variance of the sample set. Typically, when scoring entropy, we see a sharp rise in the score as we collect samples, then the score tapers off and decreases. This decrease occurs as we inevitably sample redundant information.

There are two main ways to evaluate our entropy scores. First, one might be interested in the maximum achieved entropy score. This is the value at the highest point in the curve, and indicates the highest overall performance of each algorithm. Alternatively, one might be interested in the entropy of the sample set after the budget is exhausted. This is the final point in our graphs, and is a good indicator of the relative performance of each algorithm over the entire test.

As an alternative to entropy, we also consider reconstruction error. Reconstruction error is calculated as the average non-negative least-squares error when using our collected samples to reconstruct all unobserved samples. Given a set of collected samples $X = \{x_1, ..., x_n\}$ and unsampled points $Y = \{y_1, ..., y_m\}$, we want to find the set of weights $W = \{w_1, ..., w_n\}$ that minimize:

$$W^* = \arg\min_W ||W^T X - Y||_2 \tag{3.75}$$

$$w_i > 0, \ \ \forall\, w_i \in W \tag{3.76}$$

This metric rewards a sample set that is representative of the dataset. While none of our test methods have reward functions that explicitly consider reconstruction error, it is a desirable quality for a planetary robotics mission in which scientists want to collect a sample set that is representative of the underlying data.

We choose to force our weights to be non-negative because intuitively it makes sense in our application. If we consider our samples to be something like a set spectra, then they are a set of mineralogical signatures representative of the sample. As these signatures cannot exist in negative quantities, we restrict our reconstruction error score to only fight non-negative weight values when reconstructing the data.

An alternative scoring metric that we do not use here is that of mutual information. Mutual information, defined as

$$I(X;Y) = H(X) - H(X|Y) \tag{3.77}$$

Instead of entropy, which maximizes the information content of the samples, mutual information rewards the set of samples that provide the greatest overall reduction in uncertainty about the rest of the sample space [Krause et al., 2008,Caselton et al., 1992]. This makes mutual information an interesting information-theoretic approach to sample selection, in that it measures the overall understanding of the data similarly to reconstruction error, while still placing reward on the information content of the samples themselves. Instead of focusing on points that might be outliers, as we do in entropy, mutual information accounts for the information reduction in the rest of the space, making it slightly more robust to outlier selection.

We choose not to use it as a metric for a handful of reasons. Primarily, we feel that reconstruction error provides a similar metric for scoring samples that reduce the uncertainty in the dataset. Both metrics are rewarded when a single sample is selected that is similar to a large portion of the unsampled and unexplained sample space. While mutual information might fall somewhere on the spectrum between the more greedy entropy score and the more encompassing reconstruction error, it is still fairly similar to reconstruction error.

Additionally, mutual information is computationally challenging to compute. When used as a sample selection metric it is NP-complete. Calculating a mutual information score is computationally less demanding, but not trivial. While there are greedy approximations in certain situations that are within $(1-\frac{1}{e})$ of the optimal solution [Krause et al., 2008], we feel that the implementation and computational costs do not outweigh the benefit of adding it as a sampling or scoring metric.

3.4 Summary of Methods

We have described a number of approaches for selecting measurement locations in an adaptive sampling context. We summarize and compare our approaches here.

- **AGMM:** Adaptive Gaussian Mixture Models (AGMMs) build up a model of the sample space over time. They are computationally cheap and numerically stable when there are a low number of samples, yet flexible enough to adjust as new samples are collected. New points are selected based on their dissimilarity to the existing model, leading to a highly diverse set of measurement locations. The size of the model size is fixed beforehand. All AGMM-based methods have no measure of noise in the contextual data; they instead react to the noise observed by placing sample points in new mixtures or merging them with similar mixtures.

- **AGMM-AddClasses:** This variant of the AGMM uses Dirichlet process priors to grow the model over time, as evidence of new mixtures is collected. It has a similar sampling strategy as the AGMM model, but it does not require knowing the number of mixtures beforehand. This gives it more flexibility in how the model is grown over time. Points chosen by this approach are likely to be more diverse than those in the AGMM approach, as it will better model the feature space.

- **AGMM-P(Wrong):** This variant of the AGMM retains the fixed model size of the base approach, but uses Dirichlet process priors to estimate how well new points compare both to existing mixtures in the model, and a new, unobserved mixture. It chooses points that fit poorly with the existing model and are likely to belong to a new mixture.

- **One-Step Bayesian Optimization:** These approaches, **GP-EI, GP-PI, GP-UCB, GP-MaxMean**, and **GP-MaxVar**, build a Gaussian process to estimate the expected reward of unsampled points. Each approach uses a different acquisition function to select measurement locations, each offering some trade-off between maximizing the expected reward of a point (GP-MaxMean) and improving the model by sampling uncertain points (GP-MaxVar). They are reliant on good training points to initialize the Gaussian process, but in practice this has had only a minor impact on results.

Table 3.1 summarizes each approach and their strengths and weaknesses.

Approach	Sampling Strategy	Strengths	Weaknesses
AGMM	Samples points most dissimilar to model	High diversity in selected samples	Fixed model size
AGMM-AddClasses	Samples points most dissimilar to model	Grows model over time	No measure of uncertainty
AGMM-P(Wrong)	Samples points that fit poorly in model	Samples points that are poorly understood	Sensitive to choice of prior
One Step Bayesian Optimization (**GP-EI, GP-PI**, **GP-UCB**, etc)	Finds points that improve GP model or have a high expected reward	Resistant to noise, methods trade off between reward and uncertainty	Reliant on choice of training points

Table 3.1: A summary of our approaches.

Chapter 4

Related Works

Adaptive sampling is a broad topic that shares common themes with a number of other fields. This chapter will review some of the common themes in adaptive sampling and their overlap with other fields, as well as recent applications of adaptive sampling and alternative sampling methods.

In Section 4.1 we discuss sampling methods based on information theory. At its fundamental level, adaptive sampling is about assigning a reward to candidate sample points, given the current set of sampled points and the known information about any new samples. This reward might be based on the estimated amount of information in new samples or the amount of uncertainty in the model, and many adaptive sampling methods make use of concepts in information theory to estimate these metrics, with a high amount of success.

Adaptive sampling also shares a number of similarities with active learning. In active learning, an agent views a pool of samples and selects points one at a time relative to some reward function. Upon sampling the point the true class of the point is known, and the reward function for unsampled points is adjusted accordingly. In Section 4.2 we further discuss the premise of active learning, its similarities and differences, and recent works in active learning.

In Section 4.3 we discuss Bayesian optimization and its relationship to our own work. Bayesian optimization is not traditionally used in adaptive sampling, but their formulation of the sampling problem shares a number of similarities to our own problem, and their explicit use of model uncertainty is a desirable trait in our own tests.

Section 4.4 discusses previous uses of onboard spectroscopic analysis by rovers. Onboard spectroscopic analysis is a burgeoning field, especially with the recent advent of low-weight spectrometers that can fly on unmanned aerial vehicles (UAVs). These spectrometers provide insights into such topics as geologic classes, crop health, and canopy cover that are difficult to characterize with only color imagery.

Section 4.5 covers prior work in latent class modeling, specifically related to Dirichlet process models. These approaches span a broad spectrum of applications in machine learning, health care, and computer vision, yet all have the common theme of modeling an unknown number of latent models in the data. Dirichlet processes are a common tool used in these approaches, and an

overview of them is provided here.

Finally, in Section 4.6, we provide a comparison between related techniques and our contributions.

4.1 Information-Based Sampling

Selecting sampling sites under uncertainty is an important topic in robotics. Typically it involves evaluating candidate sampling sites against predicted models of the scene in order to find the subset of measurements that best reduces uncertainty in the scene.

Information-theoretic reward functions aim to learn information about parameters in environmental models. Typically this is done by choosing sample points that reduce the posterior model uncertainty. If this uncertainty in our parameters Θ is quantified with Shannon entropy $H(\Theta)$, then information-theoretic functions reward observations that provide the greatest expected information gain of a new observation a:

$$IG(a) = H(\Theta) - H(\Theta|a) \tag{4.1}$$

Information gain is an apt choice for a metric when dealing with choosing a single measurement location. In this case, one finds the point with the maximum information gain and samples at that point. Information gain was used in the Nomad project to decide whether or not to deploy certain instruments to analyze a prospective meteorite [Pedersen et al., 2001]. Chaloner and Verdinelli offer a comprehensive review of other information-driven approaches to sampling [Chaloner and Verdinelli, 1995].

Shewry and Wynn introduce the idea of Maximum Entropy Sampling (MES), showing that, under some weak assumptions, one can maximize the entropy of the observations themselves [Shewry and Wynn, 1987, Cressie, 1991, Sebastiani and Wynn, 2000]. MES is an information-theoretic approach that aims to find the best set of sampling points A within the collection of all points V that have the maximum joint entropy:

$$H_V(A) = \ln \det(\sigma_{A,A}) \tag{4.2}$$

where $\sigma_{A,A}$ is the subset of the covariance matrix indexed by the data points A. MES selects the set of prospective sampling sites with the greatest uncertainty, an intuitive measure for sampling unknown locations. Solving this problem is NP-hard, but there are a number of heuristics used to find approximate solutions.

4.1.1 Spatial Design

These information-based approaches have so far only considered the case of selecting a single point. If we instead want to select a set of points, such as in path planning or in the placement of a network of sensors, we instead consider the case called *spatial design*. Spatial design considers

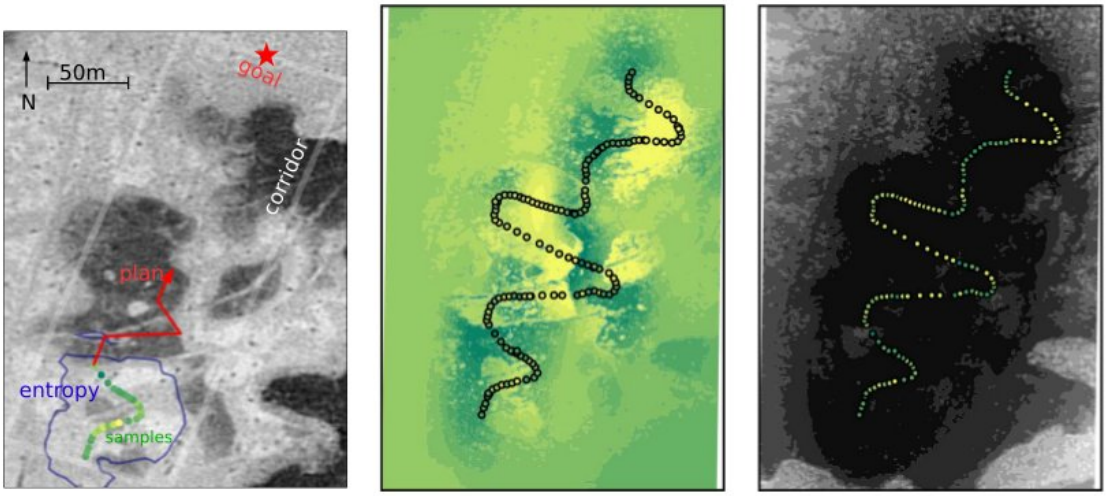


Figure 4.1: An example of the approach demonstrated in Thompson's work [Thompson, 2008]. Left: aerial imagery of a test site in Amboy Crater. Colored dots show rover spectra collected along the path. Center: geologic map showing the predicted surface material classes resulting after the traverse. Right: marginal prediction variance of the given map.

the selection of a set of points A, out of a set of possible measurement locations, V, making our model parameters $\Theta = V \setminus A$.

Shewry and Wynn apply their MES approach to spatial design, noting that their solution simply becomes the points that maximize the differential entropy of the samples:

$$R_{MAXENT}(A) = \frac{1}{2} \log \{(2\pi e)^n |\Sigma_{A,A}|\} \tag{4.3}$$

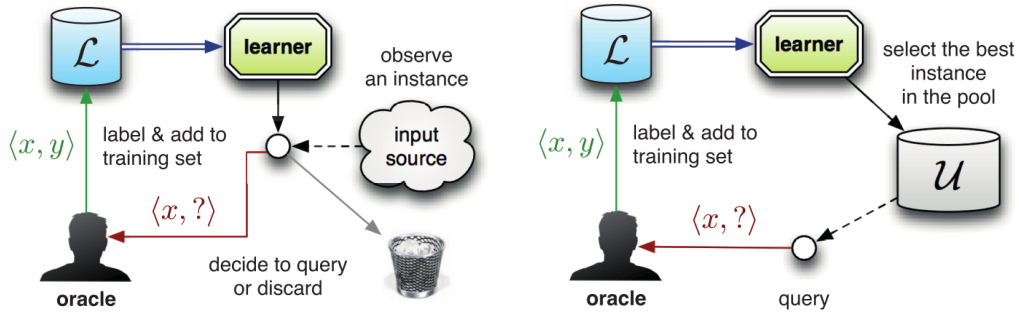
where $\Sigma_{A,A}$ is the sub-covariance matrix of the samples A.

Krause, et al. [Krause et al., 2006] propose an alternative using a mutual information criterion in their work on finding optimal sensor placements within a map. They aim to choose the best subset of sensor placement sites A within all candidate sites V that maximizes the reward function:

$$R(A) = H(V \setminus A) - H(V \setminus A|A) \tag{4.4}$$

This reward function shares similarities with MES but removes any reward from the observations themselves, causing the metric to avoid high-entropy edge locations. This approach works well in sensor site placement applications in which one wants to actively avoid placing sensors at border regions. They also show that the submodularity, the property that adding more sensors has a diminishing return, of mutual information provides good theoretic bounds on efficient greedy sensor placement algorithms.

Applications of these algorithms are extensive, so we focus on one of the most relevant ones here. Thompson explores the use of both MES and mutual information when evaluating sample



(a) Stream-based sampling

(b) Pool-based sampling

Figure 4.2: An example of stream- and pool-based sampling methods. In stream-based methods, such as our microscopic scenario, points are considered one at a time and either sampled or discarded. In the pool case, like our full context and path planning experiments, the full contextual data is available, and the agent selects a point from this pool of samples. From [Settles, 2010].

sites within a robotic traverse [Thompson, 2008], shown in Figure 4.1. His approach samples a list of candidate paths between the rover's current location and a goal, computing the entropy of the path as the joint entropy of observations along that path conditioned, on all prior samples. He uses a Gaussian process with orbital maps as a latent input to model the expected sample readings at future sites, and finds that his adaptive approach outperforms other candidate approaches.

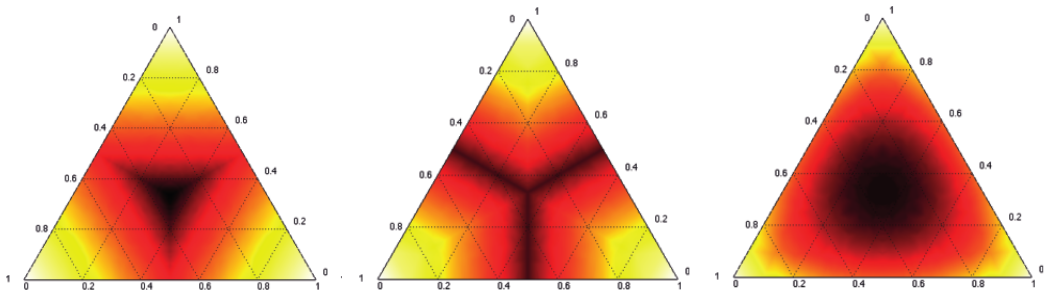
Lastly, we note the difference between spatial design and *sequential design* [Williams and Notz, 2000, Busby, 2009]. In spatial design, the experiment is designed before samples are collected. In our work we instead focus on sequential design, in which the process is iterative, and collected samples alter future rewards.

4.2 Active Learning

Active learning is a field closely related to adaptive sampling. Active learning focuses on the scenario on which data is presented to an agent, then the agent is allowed to query for more information on a single data point. The query reveals the true information about the point, such as the point's class membership, and the agent's model is updated with the new information. Active learning is often called "query learning", and is closely related to sequential design, mentioned above.

Active learning is appropriate when there is a large amount of data and an oracle that is equipped to answer the desired questions. It has been used in numerous fields, such as speech recognition [Zhu and Goldberg, 2009], information extraction from documents [Settles et al., 2008a], to computational biology [Smith et al., 2011].

There are two main approaches used in active learning, shown in Figure 4.2. The first is *stream-based sampling* [Atlas et al., 1990, Cohn, 1994], in which a single observation is available at a time, and the agent must decide whether to query that point or discard it and move on to the next point.



(a) Least confident (b) Margin (c) Entropy

Figure 4.3: An example of different uncertainty measures on a three-class toy case. Lighter colors indicate a higher probability of a sample being from nearest class label. From [Settles, 2010].

This is the case in our microscopic, limited context experiments, in which the PIXL instrument cannot return to a prior point and must decide immediately whether or not to sample at that location.

The second is *pool-based sampling* [Lewis and Gale, 1994], in which all of the data is visible to the agent, and the agent must select a sample point to query within that full set. This is the case most commonly considered in active learning research. Our baseline and orbital path planning experiments are most similar to this case in that the full data is observable, and we must select the point within this data to sample.

Within the active learning framework, there are a number of selection methods for choosing which sample points to query. One approach is that of uncertainty-based samplers [Lewis and Catlett, 1994], which are similar to the information theoretical approaches we discussed above.

Uncertainty-based approaches falls under three main categories: confidence-based sampling, margin sampling, and information theory approaches. Uncertainty-based samplers estimate a probability of class membership, then choose the least probable point. This is similar to our P(Wrong) approach outlined in Section 3.1.3. Margin-based samplers compute the margin between the two most likely predictions in the model, then sample along the boundary between the classes. Information theoretic approaches use metrics such as Shannon entropy to select points, similar Maximum Entropy Sampling [Settles, 2009]. The expected rewards of a three-class example is shown in Figure 4.3.

These uncertainty-based metrics can also be extended well to the regression case. In that situation, the agent often queries for the unsampled instance for which it has the highest prediction variance [Settles, 2010].

In addition to the metrics described above, various active learning approaches quantify the informativeness of unlabeled data using metrics such as model uncertainty [Lewis and Gale, 1994], expected model change [Settles et al., 2008b], error reduction [Roy and McCallum, 2001], or reduction in model variance [Cohn, 1994]. A good summary is provided in [Settles, 2010].

4.3 Bayesian Optimization

Research in Bayesian optimization focuses on the optimization of an unknown nonlinear function, through the sampling of some surrogate function, typically a Gaussian process [Brochu et al., 2010b]. While information theory and active learning approaches to sampling focused on sampling from a variety of reward functions, our use of Bayesian optimization instead focuses on how to take one of these established reward functions, namely differential entropy, and apply Bayesian optimization techniques to improve our sampling from within that function.

Our intuition is that purely optimizing a reward function over our contextual data, which we know to be noisy, will lead us to often sample points that are only appear interesting because of noise, but in fact do not offer the optimal reward. Instead, we turn to Bayesian optimization techniques to better estimate the uncertainty in our reward function, then explicitly make use of this uncertainty to improve our sample selection.

Bayesian optimization is incredibly useful when the underlying function is difficult to optimize or hard to sample. In fact, Bayesian optimization has been shown to be one of the most efficient approaches in terms of number of function evaluations required [Mockus, 1994, Brochu et al., 2010b, Jones et al., 1998, Sasena, 2002].

In the field of robotics it has performed very well when used to optimize cost functions used in robotic control. Tesch et al. use Bayesian optimization to learn control parameters for moving a snake robot over obstacles [Tesch et al., 2013b, Tesch et al., 2013a]. Calandra et al. compare Bayesian optimization techniques for optimizing bipedal locomotion [Calandra et al., 2014].

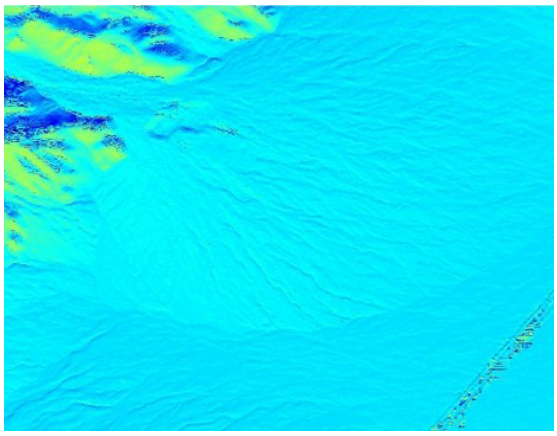
Gaussian processes and Bayesian optimization techniques have also been used extensively in the spatial modeling domain. Foil et al. use Gaussian processes to model subsurface hydrogen readings at a site in the Mojave Desert in California [Foil et al., 2016]. This work took measurements collected by a rover-mounted neutron spectrometer, used to measure subsurface hydrogen quantities, and combined those measurements with orbital data to extrapolate measurements at unsampled locations within the sample site. Measurements were improved by integrating computational models of the main physical processes that affect subsurface hydrogen transport in the Mojave Desert, namely water flow, wind flow, and insolation. Example maps from that process are shown in Figure 4.4.

In other works, Thompson uses a Gaussian process with orbital maps as a latent input to model the expected sample readings at future sites [Thompson, 2008]. Conford et al. use Gaussian processes to model wind flow discontinuities [Cornford et al., 2002].

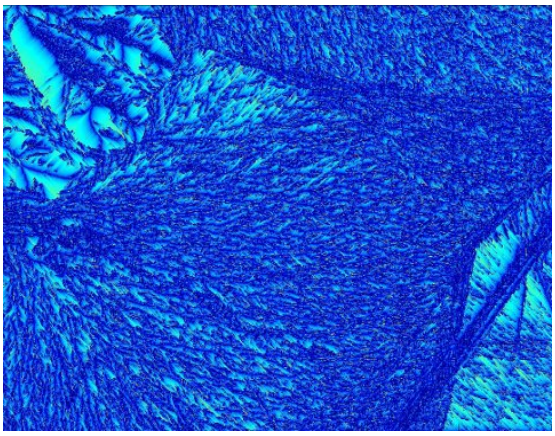
There is also some overlap between active learning and Bayesian optimization. Kapoor et al. use Gaussian processes for object categorization [Kapoor et al., 2007]. They select candidate unlabeled images so as to maximize the margin of the Gaussian process classifier and minimize the uncertainty.

Our use of Bayesian optimization techniques is different than how they are typically applied. In most experiments, the function to be estimated can be repeatedly queried, and it is then thus a problem of searching high-reward regions of the function, or regions with high uncertainty that have the possibility of having a high reward, in order to characterize the function. In our case,

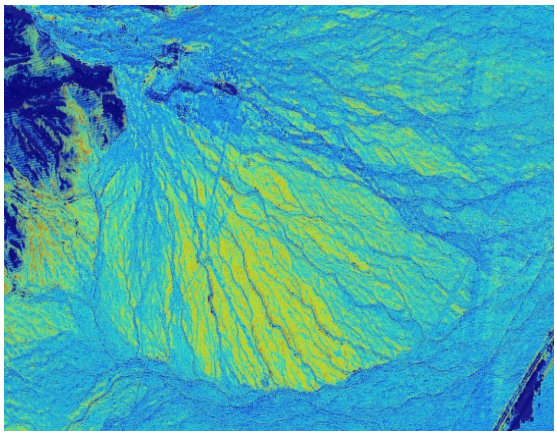
(a) Orbital view



(b) Insolation model



(c) Wind flow model



(d) Predicted subsurface water

Figure 4.4: Results of the MVP field test in which neutron spectrometer measurements estimated subsurface water distributions are input into a Gaussian process. Adding process models like insolation and water flow improves the estimate, even in the driest times of the year. The predicted map, shown on the bottom right, can be used by scientists to estimate quantities at unvisited areas of the map [Foil et al., 2016].

collecting a sample vastly changes the entropy reward at unsampled points, so we instead only sample once from the function, then recompute our Gaussian process for the next iteration.

4.4 Onboard Use of Spectra in Adaptive Sampling

With the recent development of mobile field spectrometers, a number of groups have used spectroscopic data for adaptive methods for robotic planning and sampling. One of the first approaches to use spectra onboard a rover was that of the Nomad field team [Pedersen et al., 2001]. Nomad was a

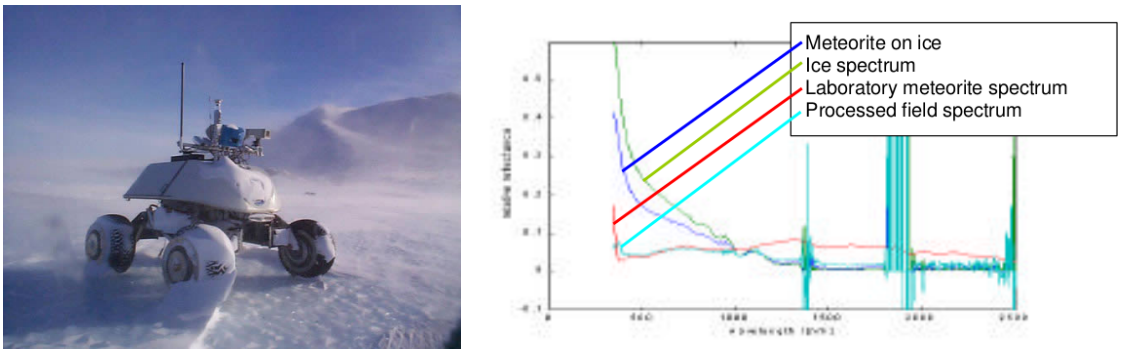


Figure 4.5: The Nomad rover, shown here, was used to detect meteorites in Antarctica. Right: Nomad used a color camera to preselect possible targets, then followed up with spectrometer readings. Left: Spectra were classified as ether meteorite or non-meteorite using a Bayes network [Pedersen et al., 2001].

rover designed to traverse regions of the Antarctic in search for meteorites. The rover was equipped with color cameras and an onboard reflectance spectrometer, used to distinguish meteorites from common rocks. The rover would detect probable rocks with its cameras, then approach the rock and take a detailed spectra, focusing on the near infrared regions, typically sufficient to distinguish rocks from meteorites. It then used prior information from spectra of known meteorite samples fed into a Bayes network to classify objects as meteorites.

Thompson et al. investigated using Gaussian process models to predict geologic classes using onboard collection of reflectance spectra [Thompson and Wettergreen, 2008, Thompson, 2008]. In his experiments he used a rover to collect high-resolution spectra periodically during a traverse in Amboy Crater, California, with the objective of predicting the presence of basalt and sediment features within a map. He tested two formulations, one in which the rover fed spectroscopic measurements into a Gaussian process with only information about the rover's location, and one in which the Gaussian process was augmented with orbital data as latent input. Results showed much higher performance with the addition of the orbital information.

Additionally, aerial spectrometers have recently received a lot of development due to the availability of low-cost UAVs. Burkart, et al. detail work in which they test a 230 gram visible to near infrared (VIS/NIR) range (350 - 800 nm) spectrometer for agriculture analysis [Burkart et al., 2014]. Aubrey, et al. discuss using gas spectrometers onboard a UAV to detect gas distributions, characterize their abundance, and track plumes [Aubrey et al., 2014]. Similarly, Diaz et al. use miniaturized mass spectrometers onboard inexpensive and lightweight fixed-wing UAVs to analyze volcanic plumes [Diaz et al., 2015].

In the microscopic sampling domain, Thompson et al. introduce a method for adaptive sampling on the PIXL instrument that uses a weighted distance metric to determine when a sample is informative and trigger a long-dwell spectrum collection at that location [Thompson et al., 2014]. This approach has the advantage of having a number of tunable parameters that the scientists can adjust for performance, but, as we will see in our later PIXL experiments, can result in poor esti-

mates of the utility of the initial spectrum before a model is built up.

Many of these approaches search for specific signals, yet in our experiments we aim to generalize to an unknown environment in which no signals are known in advance. Contextual information allows our approach to build generative models of the terrain to learn these signals or more accurately estimate reward functions for sampling within these points. An example of this, which we use as a baseline method in our orbital path planning tests, is discussed below.

4.5 Latent Class Modeling and Dirichlet Processes

If we desire a representative collection of samples [Castaño et al., 2005, Smith et al., 2005], a popular approach is to cluster the contextual data and sample from within those clusters. K-means or Gaussian mixture models (GMMs) are popular clustering approaches and have been used to preprocess the data in order to estimate class membership. These clusters can then be used for selecting representative samples in active learning [Nguyen and Smeulders, 2004], for warm-starting active learning methods [Kang et al., 2004], or for hierarchical methods that pre-cluster the data, then split the clusters with the most variance [Dasgupta et al., 2008].

In some scenarios, for example if the noise in the contextual data is low, it may be sufficient to set the number of clusters equal to the number of samples. In this case, K-means provides a local minimum with regards to reconstruction error. However, if there is noise in the system or if more complicated models are used to evaluate the expected rewards of points, then it is more difficult to use clustering methods, as the number of underlying clusters in the data is not known in advance. Some clustering techniques, such as mean shift [Vatturi and Wong, 2009], do not require a class estimate, but can be computationally intensive or sensitive to initialization parameters.

In many cases, this prior knowledge of the number of clusters is not available in advance, or may change over time. To solve this problem, the use of Dirichlet distributions and Dirichlet processes was proposed as a way to formalize class assignment under uncertainty [Ferguson, 1973]. *Dirichlet distributions* are essentially distributions over parametric distributions, meaning it is a distribution over the parameters of another distribution. For example, in the case of multinomial models, like in the case of each side of a die having a probability of being rolled, Dirichlet distributions can be used as a distribution over those probabilities. *Dirichlet processes* are generalizations of these distributions that handle an infinite number of parameters for an unknown number of classes. They are often used in situations such as ours where both the parameters of each model and number of models is unknown.

An example of this is the process proposed in Latent Dirichlet Allocation (LDA) in which there are a number of topics and the aim is to discover the latent topics each document belongs to [Blei et al., 2003]. A Dirichlet process prior is placed on the words in each topic as well as the topics of each document. The only truly observable data in the model are the words themselves, and all other information is latent.

Dirichlet processes are used in a number of other model selection applications. Wood et al. use an infinite Gaussian mixture model to model spike sorting [Wood et al., 2006]. Spike sorting is the

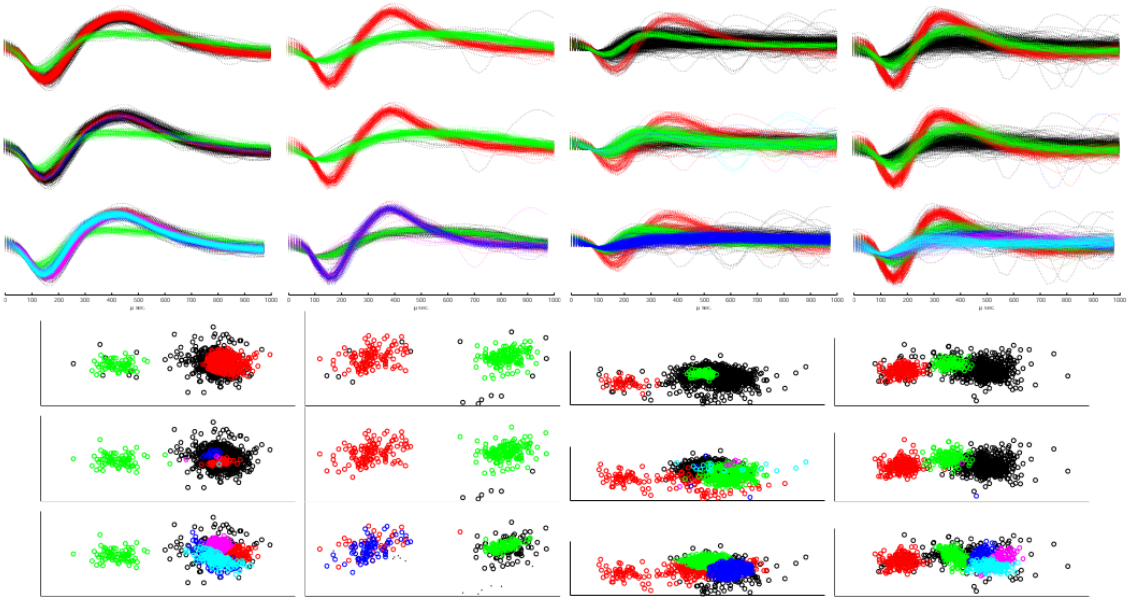


Figure 4.6: An example of Dirichlet processes being used to discover neuron counts when only a mixed neuron spike signal is available [Wood et al., 2006]. The figure shows the results of neural data when clustered using an EM approach (top rows), GMM approach (center rows), and a GMM approach using Dirichlet processes to grow dynamically (bottom rows). The number of neurons involved in each spike is unknown, but is uncovered best using the GMM augmented with a Dirichlet process prior.

process of distinguishing multiple neuron spikes by cells, illustrated in Figure 4.6. Similar to our objective, the number of cells responsible for spike signals is unknown a priori, and the objective is to correctly estimate the number of cells firing as well as their characteristic signal. Wood et al. solve this with an infinite Dirichlet process in which parameters are estimated using a Gibbs sampling approach [Neal, 2007]. Results compare favorably against a standard Gaussian mixture model tested with varying numbers of mixtures.

An interesting extension to Dirichlet processes is in the study of the hierarchical Dirichlet process in which groups of models share certain attributes [Teh et al., 2006, Fox et al., 2007, Sudderth et al., 2005]. Fox et al. describe an approach that uses a hierarchical Dirichlet process to find objects within images in which the total number of objects is unknown. As the same object in two images has similar properties, these properties are shared within images to create a more robust model. They then extend this model observing that object qualities within images share similar covariances but different means, such as the same object placed at two different pixel locations in two images. Their results compare favorably against non-hierarchical models like LDA. This work has particular relevance to our objective of retaining a history of data across scenes.

Friedman et al. describe an approach for active learning applied to Dirichlet processes [Friedman et al., 2011]. Their objective is to analyze large amounts of underwater UAV data, model an

unknown number of classes, and assign class labels. They use active learning to query a human for true class labels on uncertain samples. This has obvious parallels with our approach, in which the rover performs its own queries by gathering measurements at points of ambiguity, however they focus solely on learning class models and ignore our alternate objective of novelty detection.

4.5.1 Dirichlet Priors on Class Creation

Perhaps the example most similar to our approach is that of Haines et al. [Haines and Tao Xiang, 2013], discussed briefly in Section 3.1.3. Haines considers a pool-based active learning case which aims to jointly discover rare classes from within a dataset and refine the parameters of a classifier over the unsampled data. Like many active learning approaches, the data consists of sample points, each of which has some feature descriptor as well as a true, unobserved class label. Sampling at a location reveals the true class, but the number of samples collected is constrained with regards to a budget on the number of total sample observations.

Their main contribution is in the use of Dirichlet processes as a nonparametric estimator on the probability of any sample coming from an unknown and, as yet, unobserved class. His approach consists of three steps. First, each potential measurement location is evaluated with regards to the existing classification model, and a probability of each point x_i belonging to each model, $c \in C$, is calculated:

$$P_c(x_i|c),\ \forall c \in C \tag{4.5}$$

This probability is then modified to include the probability of a sample potentially belonging to an unobserved, unknown class, estimated using a Dirichlet process:

$$P_n(c \in \{C \cup new\}|x_i) \propto \begin{cases} \frac{m_c}{n-1+\alpha} P_c(x_i|c) & \text{if } c \in C \\ \frac{\alpha}{n-1+\alpha} P(x_i) & \text{if } c = \text{new} \end{cases} \tag{4.6}$$

Second, the probability of misclassification is calculated for all known models, accounting for the potential misclassification of incorrectly identifying that the sample is from an unknown class:

$$P(wrong|x) = 1 - P_n(c'|x) \tag{4.7}$$

Finally, the point with the greatest probability of misclassification, c', is selected and sampled, and the underlying class is revealed:

$$c' = \arg\max_{c \in C} P_c(x|c) \tag{4.8}$$

In their work, the goal is to balance finding new classes with refining existing ones. Many approaches have done this by finding points that have high uncertainty with regards to the existing models, but most, such as entropy, cannot explicitly account for unknown classes. Other approaches, such as sampling at the points with the highest probability of misclassification with

regards to only existing class models, have done well at refining models, but do not account for unknown models without explicit consideration.

Haines' goals mirror those of this dissertation in some ways, specifically the desire to find rare classes and refine existing models. However, while Haines considers a classic classification approach, in which each point of the data is assigned a specific class label, our work has no such labels, and must instead estimate class distributions only from prior observations. In practice, this has led to complications in the estimation of class membership probabilities, as well as a number of other minor effects on the data. Furthermore, the use of a Dirichlet prior on unknown classes is reliant on an estimate of the probability of any given data point, which, if it can be calculated at all a priori, is often very low in the high-dimensional data we work with. This leads to a minimization of the contribution of the Dirichlet process prior, essentially turning the approach into an estimate of misclassification only under known classes, a scenario Haines et al. were specifically trying to improve upon.

4.6 Comparison of Methods

In this dissertation we consider methods that use noisy contextual data to estimate the best locations to sample. The related works described here have approached similar problems in a number of different ways, and each of them has their strengths and weaknesses when used in adaptive sampling.

However, many of these approaches do not extend well when there is a large amount of noise in the contextual data. Uncertainty-based sampling approaches, such as Maximum Entropy Sampling, are prone to poor sampling due to noise. Without any explicit noise model or some uncertainty in their own rewards, these methods often sample points that are only interesting because they are corrupted with high noise.

Active learning methods suffer from many of the same problems as uncertainty-based sampling, as they share similar reward functions. Additionally, active learning commonly considers the case of classification, in which samples have class labels, which simplifies much of the problem as it allows for more accurate model generation and reward prediction. In the regression case, active learning offers little more than that offered by uncertainty-based sampling methods.

Bayesian optimization methods offer a more structured, model-based approach to sampling, yet, to our knowledge, this is the first work that has used them for reward prediction in a sequential design scenario. Commonly they are used to repeatedly sample the same unknown function, but here we use similar techniques over a reward function that changes at each iteration.

Finally, our work in Dirichlet-based approaches using modified Gaussian mixture models gives us a model-based approach with the flexibility to change over time. This results in more accurate models that do not rely on preconceived assumptions about the data. Table 4.1 compares our approaches with related works outlined in this chapter.

	Model-based	Models grow over time	Estimates reward uncertainty	Predicts expected reward
Uncertainty-Based Sampling	No	N/A	No	Yes
Active Learning	No	N/A	No	Yes
Bayesian Optimization	Yes	No	Yes	Yes
SpectroMapper	No	N/A	No	Yes
AGMM	**Yes**	**Yes**	**No**	**Yes**
Our Bayesian Optimization	**Yes**	**No**	**Yes**	**Yes**

Table 4.1: A feature comparison between our methods, in bold, and those of related approaches.

Chapter 5

Adaptive Sampling with Full Context

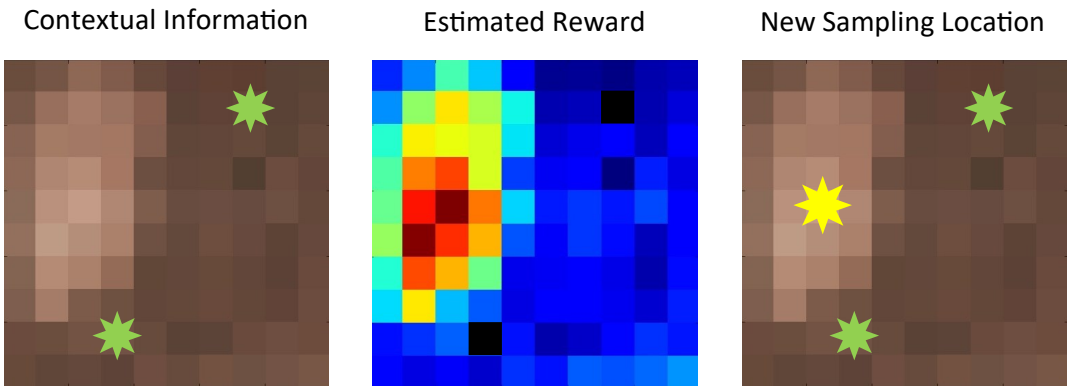
At the most basic level, adaptive sampling is about reacting to prior sample selections and improving existing models of the data. Our three experimental scenarios, shown in Table 5.1 will consider a number of sampling strategies, some well-established in literature, and others novel to adaptive sampling, with the common theme of improving our ability to both estimate the reward of measurement locations and discover rare sample signatures within contextual data.

Experiment	Contextual Data	Goal	Robotic Application
Full Context	Noisy estimate of every point	Select most rewarding sample sites	Rover selecting sample targets using navcam imagery
Limited Context	Noisy estimate of a single point	Determine when best to fully sample points or move on	Microscopic sampling using an instrument like PIXL
Path Planning	Orbital image	Plan path that samples at the most rewarding locations	Sample site selection across a rover's traverse

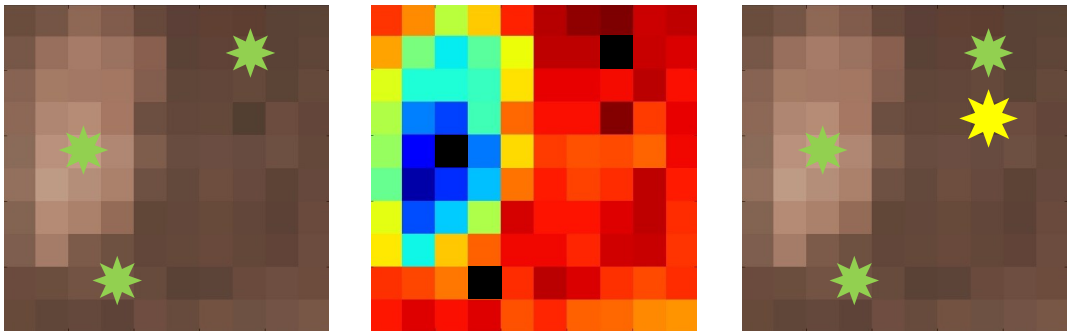
Table 5.1: A summary of the three experiments considered in this work and their analogues in robotic exploration.

The first experiment is that when the robot has access to contextual information at all possible measurement locations. This is similar in concept to a rover collecting a navcam image, then choosing targets of interest for follow-up measurements with a targeted instrument, such as a spectrometer. Because the full contextual data is available, this experiment is designed to demonstrate and compare the effectiveness of a number of approaches under a minimal number of sampling constraints.

The objective of this first experiment is to choose a sequence of points to fully observe which maximizes the understanding of the underlying mineralogy. As discussed in more detail in Section



(a) Choosing the third sample point



(b) Choosing the fourth sample point

Figure 5.1: An example of the sampling process in the general case. Left column: The full context is observed. Center: Rewards are estimated given prior samples. Right: A new measurement location is chosen that maximizes the expected reward.

3.3, the sample set is evaluated with regard to two main scoring functions: differential entropy and reconstruction error. Each of these scoring metrics provide a different balance between selecting the most unusual and different samples and selecting samples which provide the greatest degree of understanding about the unsampled points in the scene. We will draw observations on the results of each method with regards to the scoring function, but we leave it up to the reader to decide which sampling method is most relevant to any potential related work of their own.

Scores are evaluated on the raw data from the instrument and a reduced version of the data that is transformed from energies to pseudointensities. These pseudointensities approximate the abundance of specific elements in the sample, such as iron, aluminum, or sulfur. In a planetary exploration scenario the most important scores are likely those that relate directly to these pseudointensities, as they are processed to be clean of underlying effects. However, we include the

results for the full, raw dataset here as that may be applicable for a broader audience.

Finally, we discuss the results framed in a planetary exploration context. Instead of evaluating each approach purely on the numerical value of each sample collected, we look at the ability of approaches to find examples of rare elements within the data. These elemental signatures are of value to planetary geologists evaluating these samples and provide insight into how well these approaches would perform in a planetary robotics application.

5.1 Microscopic Data from the PIXL Instrument

This experiment is tested using data from the Planetary Instrument for X-Ray Lithochemistry (PIXL), a spectrometer developed at the Jet Propulsion Laboratory for detailed spectroscopic analysis of microscopic samples. The instrument is designed to be mounted on a robotic arm and analyze samples from less than five centimeters away, with a spatial resolution of one millimeter. It is currently in development and testing and is one of the seven instruments selected for the upcoming Mars 2020 mission.

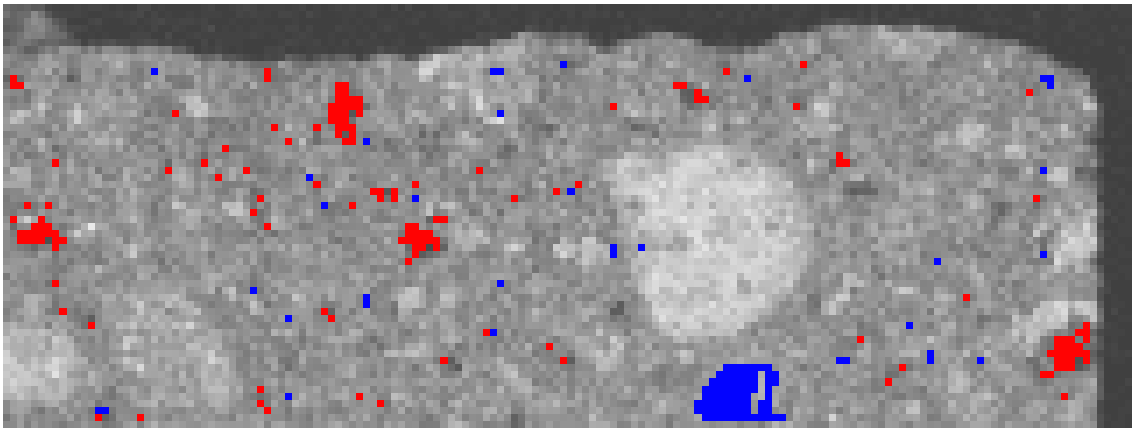
PIXL emits X-rays through a polycapillary which interact with sample material. As the X-ray strikes the sample, the high-energy, short-wavelength radiation can cause the material become ionized. If this energy is sufficient, it can dislodge an atom's inner electrons. Electrons in the outer shell of the atom then move to replace the missing inner electron, causing the emission of fluorescent radiation. The wavelength of this radiation is specific to the emitting minerals, allowing for the detection of elemental abundances in sample measurements. The instrument is equipped with a source-ray focusing optic with very high spatial resolution, allowing it to collect accurate fine scale maps of a sample surface.

The raw data collected at a single point by the PIXL instrument is a 4096-element vector that describes the electron response from 0 to 41 keV. Peaks in the data indicate abundances characteristic of minerals that commonly emit radiation in those bands. An example of the type of data collected and mineral response is shown in Figure 5.2.

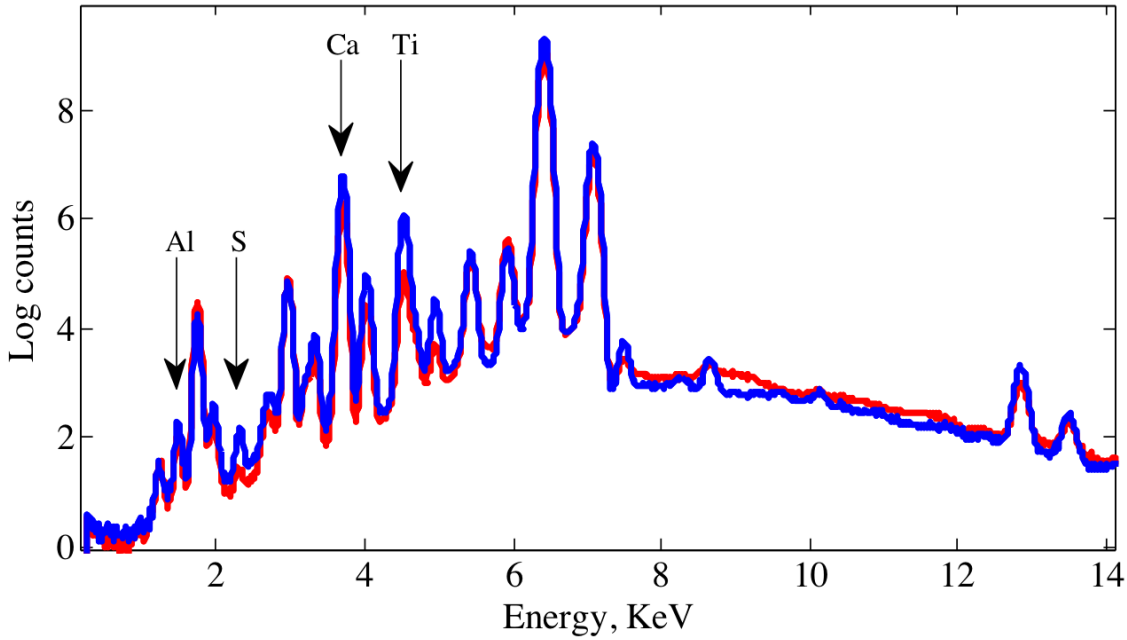
The data used in this thesis was collected using the PIXL instrument during the summer of 2015. The sample used here is from the Dressermats dataset, a sample specimen analyzed in detail by the PIXL instrument at the Jet Propulsion Laboratory. The full sample is a raster of 180 by 170 points, in which each point is a 4096-element spectrum. Each spectrum was collected using a full-duration scan, meaning the instrument integrated data over a longer period of time in order to increase the signal-to-noise ratio (SNR).

It is important to discuss the impact of integration time on the samples. X-ray fluorescence is described by a Poisson process. If N photons are returned at a specific wavelength, we assume N is a Poisson random variable with mean $N_0 = N$. The probability mass function of such a Poisson process is

$$P(N = n) = \frac{1}{n!} e^{N_0} N_0^n, \quad n = 0, 1, 2... \tag{5.1}$$



(a) “Black Beauty” PIXL sample



(b) Mean spectrum of blue and red regions

Figure 5.2: Left: Image of the “Black Beauty” sample, in which red and blue indicate materials of two different distinctive compositions. Right: The mean spectrum of the red and blue regions, respectively. Peaks indicate strong responses from certain minerals, as indicated at points in the graph. From [Thompson et al., 2015a].

with $E[N] = \text{Var}(N) = N_0$. Observing this, we can see that the absolute variance of N increases with N_0. However, if we consider the coefficient of variation (c.o.v.):

$$\text{c.o.v.} = \frac{\sigma_n}{E[N]} = \frac{1}{N_0} \tag{5.2}$$

we can see that the relative variability decreases as N_0 increases, meaning a longer duration integration gives a higher SNR. In fact, the equation for SNR involves the inverse of the coefficient of variation:

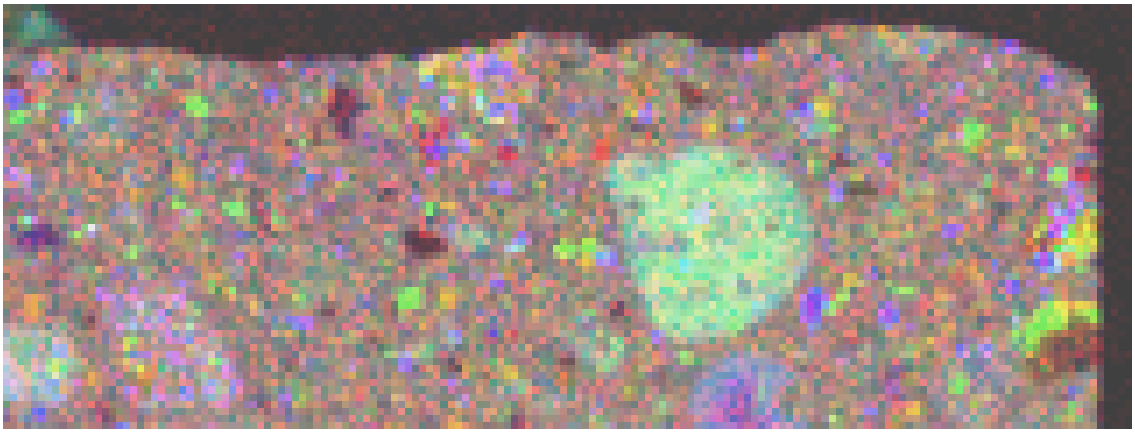
$$\text{SNR} = C\frac{E[N]}{\sigma_N} \tag{5.3}$$

As with all instruments operating under similar models, there is a trade-off between long-duration, accurate sample collection, and more practical short duration scans. In the case of PIXL, there are two sampling modes: a quick integration with low SNR, and a longer dwell integration with higher SNR. Individual measurements take anywhere from 15 to 60 seconds, so a sample of an entire surface takes an hour or two, while the more accurate, long-dwell sample takes many tens of hours. It is thus infeasible in a planetary mission, such as the Mars 2020 mission, to do a full-duration scan at every point on the surface. Instead, we propose to use adaptive sampling to analyze a set of fast, inaccurate samples, then choose a subset of the the most promising points from within the sample at which to collect long-dwell spectrum.

5.1.1 Transformation from X-Ray Responses to Mineralogy

In order to extract specific elemental abundances from the X-ray signal, we perform a number of steps to estimate the underlying signal and local background. First, we estimate the Bremsstrahlung background, shown as the smooth underlying curve to the data in Figure 5.3. This is done using the SNIP algorithm of Ryan et al. [Ryan et al., 1988, Van Grieken and Markowicz, 2001], a non-parametric estimator of the background signal. We subtract off the background signal, and in regions of low energies where the signal calculation is unstable, specifically below 3 keV, we use the Bremsstrahlung signal at 3 keV as a proxy. Next, we integrate over specific sets of wavelength ranges known to be associated with certain elements, then normalize this by its local background. These wavelength ranges are the energy channels associated with the dominant Kα1 peaks for a number of elements, such as Al, Si, Ca, S, P, Cl, and Ti. Details on the elemental energies can be found in [Bearden, 1967, Krause and Oliver, 1979], while details on the extraction of elemental abundances can be found in [Thompson et al., 2015a]

The resulting data is no longer in terms of energy, but is now a set of line intensities, which roughly correspond to elemental abundances. The full process produces abundances for 90 elemental signatures. In order to reduce the complexity of the experimental computations, we analyzed the abundances produced from the full, long-dwell dataset, and removed any elemental signatures that were not present at significant levels. This reduces the full set of signatures observed within



(a) "Black Beauty" false-color image

Figure 5.3: The "Black Beauty" sample shown in false color using the extracted mineralogy described in Section 5.1.1. The image shows the normalized intensity of elements Al (red), Ca (green), and Ti (blue). From [Thompson et al., 2015a].

the PIXL data down to 35, shown in Table 5.2. In some cases, as we will see later, reducing the data even further, to the nine channels with highest variance, provides an even larger increase in sample diversity.

Channels with observed signal	Zr, Ru, K, P, Cr, C, Ca, Y, Cu, Sr, Cl, Ga, Co, Zn, Br, Nb, Ni, Tc, Mn, Si, Rh, Ar, Mo, As, Mg, S, V, Fe, Ge, Ti, Rb, Kr, Se, Sc, Al
Channels with highest signal	K, Ca, Sr, Co, Si, Ar, S, Fe, Ti

Table 5.2: The elements with a noticeable signal and the highest amount of signal in the PIXL dataset.

5.2 Adaptive Sampling Experiments for Fully Observed Data

In this initial case, where each algorithm has access to all contextual data and must choose measurement locations, all of the approaches outlined in Chapter 3 are evaluated:

A Adaptive GMM

(a) **AGMM**: Fixed number of mixtures

(b) **AGMM-AddClasses**: Dirichlet process prior on adding classes over time

(c) **AGMM-P(Wrong)**: Probability of misclassifying new samples, given the model

B Bayesian Optimization

(a) **GP-PI**: Probability of Improvement

(b) **GP-EI**: Expectation of Improvement

(c) **GP-UCB**: Upper Confidence Bound

(d) **GP-MaxMean**: Max Mean

(e) **GP-MaxVar**: Max Variance

We compare the results against a number of established techniques, described in more detail later in Section 5.2.3.

5.2.1 Adaptive Gaussian Mixture Models

The first approach, **AGMM**, builds up distributions describing each mixture in the model as sample points are collected. The obvious method for selecting new sample points is to find those most poorly described by the existing model. The model is built up as samples are collected, then new points are selected which have the largest minimum Mahalanobis distance:

$$\arg\max_{x_{n+1}} \min_{i} \sqrt{(x_{n+1} - \mu_i)^T \Sigma_i^{-1} (x_{n+1} - \mu_i)} \tag{5.4}$$

This selects the point that is the furthest from its nearest mixture, under Mahalanobis distance defined by the existing mixtures, or essentially the point most dissimilar from its nearest mixture.

As shown in the results in Section 5.3.2, this model performs very well given the problem formulation. However, we note a handful of limitations with the approach that motivate the inclusion of Dirichlet processes described in the next sections. First, the model is limited by a pre-defined number of mixtures. Given the limited number of samples allowed by the pre-defined sampling budget, this is not too much of a problem in this basic test case, but in later formulations, specifically the microscopic case, in which no contextual data is available and samples arrive one at a time, the constraint on number of mixtures becomes more prohibitive. More detail on this will be given in Section 6.2.

Second, the model parameters in this initial formulation are very simplified, specifically the estimation of σ^*. In reality, the variance cannot be simplified to a single value across all channels unless the data is normalized to $\mathcal{N}(0, 1)$. Instead, we use the Poisson statistics of the sample to initialize the variance of each channel. Specifically, we know the PIXL data is described by a Poisson process, and that, given a sample from a Poisson process, λ_i, the variance is equal to $\mu = \sigma^2 = \lambda_i$. Using this, and working under the assumption that all channels of the data are independent, we can initialize the diagonal of the covariance to simply the value of the observed signal. When merging mixtures, we assume that the data can be described by a Gaussian centered at the mean of the merged points, with a diagonal covariance matrix with diagonal values equal to the variance of the merged points.

Adaptively Adding Classes Using Dirichlet Process Priors

Next the **AGMM-AddClasses** approach is considered, in which the AGMM approach is improved by placing a Dirichlet process prior over adding new classes. Given an estimate of α and the prior $P(x)$, the probability of a sample point coming from a new mixture can be calculated. This is done by first selecting a point, x^*, using the Mahalanobis distance metric described in Equation 5.4. Given this point, the probability of it being sampled from all existing mixtures is calculated, $P_c(x|c)$. Then, given the current estimate of α and $P(x)$, the weighted probability of misclassification, $P(Wrong)$, described in Section 3.1.3, is used to select new samples:

$$P(Wrong) = 1 - \frac{\arg\max_c P_c(x|c)}{P_{DP}(x)} \tag{5.5}$$

$$P_{DP}(x) = \frac{\alpha}{n-1+\alpha}P(x) + \sum_{c \in C} \frac{m_c}{n-1+\alpha}P_c(x|c) \tag{5.6}$$

If $P(Wrong)$ is greater than a threshold, a new mixture is added and all relevant statistics are updated. If $P(Wrong)$ is lower than this threshold, the point is added to the existing mixture that is most similar. A threshold of 0.997 was used in all tests.

In this manner, point are still selected that are most dissimilar to the existing model, just as in the **AGMM** approach, but the consideration for adding each sampled point to a new mixture is held off until the sample is collected and the true data is observed.

As discussed in Section 3.1.2, the value of $P(x)$ is crucial to the performance of the algorithm. If it is incorrectly estimated as too large, it drowns out the contributions of $P_c(x|c)$, and if it is too small it offers no contribution at all.

Because the test data for this experiment is high-dimensional and we are interested in points that have a low probability of occurrence, a uniform prior on $P(x)$ is used. This prior then acts as scaling on the Dirichlet component of the mixture creation process described above, ensuring that the scale of the Dirichlet component is comparable to the scale of the $P_c(x|c)$ component. If these two components are not comparable, then one outweighs the other, leading to either too much of a tendency to add new mixtures if $P(x)$ is too large, or a tendency to never add a mixture if $P(x)$ is too small. To make this uniform prior easily calculable, the contextual data is normalized to have zero mean and unit variance, giving a uniform prior of

$$P(x) = 0.25^d \tag{5.7}$$

This assumes a uniform probability over the range $[-2, 2]$, scaled by the dimension of the data, d.

While this may seem like a gross simplification as we know the data is drawn from a Poisson process, we note that Poisson distributions with a mean greater than $\lambda = 10$, which all of the most informative channels have, can be closely approximated by a Gaussian. Thus, the only channels that lose some information in the normalization process are those that contain little actionable information already. Lastly, this normalization process means that channels with high amounts of information before normalization are now scaled equally to those channels that had little information before normalization. To handle this, we consider only a subset of all channels in the selection

process, limiting ourselves to the channels with the most information before normalization. These channels in the PIXL data are the elements shown in Table 5.2. This process of reducing the number of channels of interest is similar to what is typically done in planetary missions today, as scientists usually have some prior knowledge of the elements of interest in a specific sample observation, and want to focus their attention on irregularities within these specific elements.

Misclassification Estimation Using Dirichlet Priors

Finally the **AGMM-P(Wrong)** approach is evaluated, which changes how points are selected compared to the **AGMM** method. Instead of a Mahalanobis distance metric, the probability of misclassification, $P(Wrong)$, is used to score unsampled locations. This probability is defined as:

$$P(Wrong|x) = 1 - P_n(c'|x) \tag{5.8}$$

$$c' = \arg\max_{c \in C} P_c(x|c) \tag{5.9}$$

where P_n is the Dirichlet-weighted probability of belonging to an existing class or a new, unobserved class:

$$P_n(c \in \{C \cup new\}|x) \propto \begin{cases} \frac{m_c}{n-1+\alpha} P_c(x|c) & \text{if } c \in C \\ \frac{\alpha}{n-1+\alpha} P(x) & \text{if } c = \text{new} \end{cases} \tag{5.10}$$

At each step, the point with the greatest value of $P(Wrong)$ is chosen for sampling and is added to the AGMM using the original insertion and merging rules defined in Section 3.1. Our experiments use a fixed number of 20 classes and do not add more classes beyond that value.

Similarly to the approach defined in Section 5.2.1, a uniform prior on $P(x)$ is assumed and the data is normalized to be zero mean and unit variance.

5.2.2 Bayesian Optimization

Additionally, the Bayesian optimization strategies discussed in Section 3.2 are evaluated. In this scenario, we build a Gaussian process over the expected entropy score for all candidate sample points. As discussed in Section 3.2, the Gaussian process is trained on the expected entropy score of a random subset of 400 unsampled points.

Shown in Table 5.3 are the acquisition functions considered in this work. These are chosen as they offer a varied selection of approaches and have been shown to perform well in other domains.

As recommended by Lizotte [Lizotte, 2008], the PI and EI functions use a ξ parameter of $\xi = 0.01$ if the data is normalized to draw from $\mathcal{N}(0, 1)$. If the data is not normalized, as indicated in some of the tests, ξ is scaled by the variance to a comparable value.

Finally, some of the acquisition functions require an incumbent value, x^+. This is traditionally the highest observed value of the objective function over previous iterations. As the objective function changes as new samples are acquired, instead the incumbent is defined as $x^+ = \arg\max_{x_i \in X_{unsampled}} \mu(x_i)$, or the point with the highest observed maximum value.

GP-PI - Probability of Improvement [Kushner and Yin, 1997]	$PI(x) = \Phi\left(\frac{\mu(x)-f(x^+)-\xi}{\sigma(x)}\right)$
GP-EI - Expectation of Improvement [Mockus et al., 1978] [Jones et al., 1998]	$EI(x) = \begin{cases} (\mu(x) - f(x^+) - \xi)\Phi(Z) + \sigma(x)\phi(Z) & \text{if } \sigma > 0 \\ 0 & \text{if } \sigma = 0 \end{cases}$
GP-UCB - Upper Confidence Bound [Srinivas et al., 2010]	$UCB(x) = \mu(x) + \sqrt{\nu\tau_t}\sigma(x)$
GP-MaxVar - Maximum Variance [Lewis and Gale, 1994]	$\sigma_{max} = \arg\max_x \sigma(x)$
GP-MaxMean - Maximum Mean	$\mu_{max}(x) = \arg\max_x \mu(x)$

Table 5.3: Tested acquisition functions

5.2.3 Baseline Methods

Our approaches are compared to three established techniques. First, they are compared against a greedy entropy maximization approach, shown in tests as **Max Entropy**. This approach calculates the expected differential entropy for each point, given the existing sample set:

$$x_{n+1} = \arg\max_{x_i} H(X \cup x_i), \ \ \forall x_i \notin X \tag{5.11}$$

where X is the existing set of fully-observed samples. This is a fairly simple approach but performs quite well, especially in scenes with low noise. If the experimental goal is to maximize the differential entropy of the samples, as discussed in Section 3.3, then, under no noise, this method achieves the highest possible score. However, given more and more noise in the system, this approach begins to perform suboptimally and is eclipsed by alternative methods.

Next, results are compared to a grid-based, or periodic sampling strategy, shown in tests as **Grid**. In this case, the budgeted sample points are uniformly spread across the sample space. Grid-based strategies require no onboard analysis by the robot, and provide an even, unbiased representation of the environment.

Additionally, results are compared to random sampling, shown in tests as **Random**. As expected, random sampling chooses a point to sample at random within the contextual data. This is truly the baseline method, and we expect all approaches to improve upon this.

Finally, we consider an approach, called **Integral Entropy** in tests, that directly makes use of knowledge about the noise of the instrument used to collect this data. As the data is known to come from a Poisson process, the expected value and variance in the value of each point λ_i is equal to

the value itself: $\mu = \sigma^2 = \lambda_i$. Earlier this fact was used to create sample test datasets that were noisy representations of the true, underlying data. Here, it is instead used to integrate over the potential range of scores for a given sample point, to create a better estimate of the true reward at that location.

To do this, Monte Carlo integration is used. Monte Carlo integration estimates complicated integrals by approximating the integral using a large number of random samples drawn from the original function. In this case, the complicated integral is:

$$E[H(X \cup \boldsymbol{\lambda})] = \int_0^\infty ... \int_0^\infty P(\boldsymbol{\lambda}) H(X \cup \boldsymbol{\lambda}) d\lambda_1 ... d\lambda_n \tag{5.12}$$

where $\boldsymbol{\lambda}$ is the noisy observed value, X are the points that were already sampled, and the integral is done over all elemental channels. This is computationally challenging, so instead Monte Carlo integration is used to approximate this value:

$$E[H^*(X \cup \boldsymbol{\lambda})] = \sum_{i=0}^{n} P(\boldsymbol{\lambda}^*) H(X \cup \boldsymbol{\lambda}^*) \tag{5.13}$$

$$\boldsymbol{\lambda}^* \sim \text{Poisson}(\boldsymbol{\lambda}) \tag{5.14}$$

for some large value of n. The sample point selected is then the point with the largest expected entropy score:

$$x_{n+1} = \arg\max_{x_i} E[H^*(X \cup x_i)] \tag{5.15}$$

where $\boldsymbol{\lambda} = x_i$.

This results in a smoothing of the data similar to what is seen in the Bayesian optimization approaches, and a less greedy approach to sampling as the Max Entropy approach above. Scores are less impacted by the individual values of the data, resulting in a better estimate of reward.

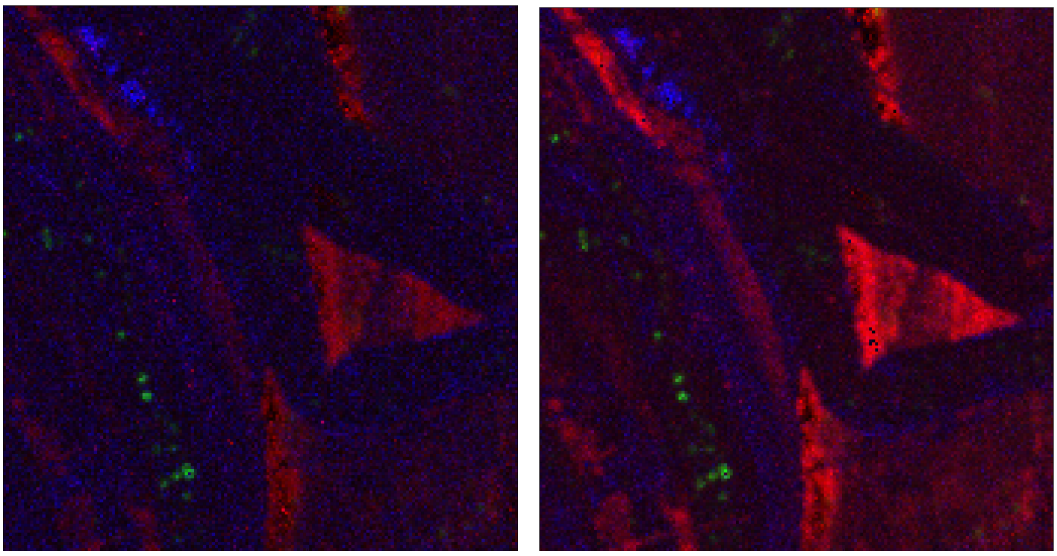
5.2.4 Testing Methodology

To test each approach a number of test contextual images are synthesized. This is done by treating the original data as a clean, fully-observed samples. Knowing that the PIXL data is drawn from a Poisson process, the value of each element at each point can be treated as the expected value of the Poisson distribution for that element at that point:

$$\lambda_{i,b} = \text{PIXL}(x_i, b) \tag{5.16}$$

$$\tag{5.17}$$

in which b is the channel that describes the elemental response at this point, and x_i is a location within the image, and $\lambda_{i,b}$ is the Poisson random variable associated with this location. To create



(a) Original PIXL image (b) PIXL image with added Poisson noise

Figure 5.4: An example of the augmented PIXL training set. On the left is a false-color image showing the abundances of cobalt (red), calcium (green), and strontium (blue). On the right is an image of the original PIXL data augmented with Poisson noise, showing the same three elements. The overall structure of the image stays the same, but the relative abundances of each element are slightly different at each pixel.

a noisy version of the data at this location and for this element, a sample is drawn from a Poisson distribution using the expected value, $\lambda_{i,b}$:

$$\text{Noisy}(x_i, b) \sim \text{Poisson}(\lambda_{i,b}) \tag{5.18}$$

In this manner multiple contextual images are created, each based on the Poisson statistics of the original data. This data is then converted to pseudointensities per the method outlined in Section 5.1.1.

In this experiment, 100 versions of this contextual data was created and each approach was tested once per image. An example of an output of this process is shown in Figure 5.4.

All approaches were initialized by fully observing the top left point in the image, then proceeded to select 24 more sampling points at any location within the contextual data, chosen by maximizing each of their reward functions.

When a sample point is chosen, that location is considered observed, and the original, clean data is added to the sample set. This sample set is used by many of the approaches to estimate the reward of future points, or to build up models of the data used for later sample selection. The scoring metrics, described in Section 3.3, are based on the values of the data in this set of clean samples. The final results are averaged across all tests.

5.3 Results and Discussion

Here we discuss the results from each method and the parameters used to generate each of the indicated tests. We begin by showing the results of the baseline methods, grid sampling, random sampling, and greedy entropy maximization. We then compare these baseline methods against each of our contributions: the AGMM and Dirichlet-based approaches and approaches using Bayesian Optimization. We finish with a comparison of the best-performing approaches in each category.

We find that some approaches perform better on normalized or non-normalized data, and some show a marked difference when using all 35 element channels compared to only the nine channels with the highest variance. In the results we indicate if an approach is using normalized data ($D \sim \mathcal{N}(0, 1)$) with the note "**(Norm)**" in the method name.

Finally, we compare results using scores from both the raw, full-channel data, with over 4000 spectral bands, and the reduced-channel data, with 35 elements. We note that in planetary robotics the reduced dataset is likely more relevant, as post-processing will remove some known noise and underlying trends in the data, like the Bremsstrahlung signal discussed in Section 5.1.1. However, the results on the full data are also included as that may prove useful for some readers and offers slightly alternative results.

The results here are fairly exhaustive, covering approaches, data processing, and scoring metrics. The aim of this section is to show under which conditions these methods perform well, then select the top-performing methods for later tests. Section 5.3.4 will show the relative performance between methods and discuss which methods are best applied to the later experiments.

The graphs plot the entropy change over time for each approach, in which a higher score is better, as well as the maximum entropy achieved and the final entropy after 25 sets. This value is expected to reach some optimum then decline over time, as we eventually are sampling repeated data and the variance of the sample set decreases.

Additionally, we show the reconstruction error, in which a lower number is better, as well as the final reconstruction error after all samples. All results are averaged over 100 trials.

We end with an evaluation of the elemental signatures detected by each approach. While having a high entropy score or low reconstruction error is good, approaches must also generalize well to a planetary robotics mission. This section evaluates each approach on the number and rarity of elemental signatures fully sampled by each approach.

5.3.1 Baseline Results

The results for the baseline approaches are as expected. Both **Random** and **Grid** provide a lower baseline for comparisons, while **Max Entropy** and **Integral Entropy** provide a marked improvement over the two simple methods. Results are shown in Figure 5.5 and Figure 5.6. The former demonstrates the results on the full-channel data, while the latter demonstrates results on the reduced-channel data.

In both tests, the **Max Entropy (Norm)** method that is running on the normalized data has slightly higher performance than the **Max Entropy** method without normalization, although the

differences are not statistically significant. Normalization provides a slightly better scores in the first five to eight samples, but both result in very similar overall scores.

Integral Entropy gives an interesting result. Because it smooths over some of the noise in the observed data, it is expected to offer an improvement over the **Max Entropy** method, and a slight improvement can be seen in the entropy of the first ten samples of the full-channel test. However, after ten samples, its entropy score slips below that of **Max Entropy**, and its final scores are weaker. This might be because the smoothing done by the integration causes the method to select diverse, safe points initially that would be overlooked by the greedier **Max Entropy** method, giving it a strong start in scores. However, the smoothing potentially limits the ability of the method to locate really unique signatures, resulting in an overall decrease in performance across the entire sample set.

Under the pseudointensity tests, **Integral Entropy** performs worse than **Max Entropy** approaches, with regards to both entropy and reconstruction error. This is interesting, as the integral was designed to account for noise model specifically tailored for this data. It is possible that converting the data to pseudointensities decouples the data somewhat from a true Poisson process, or perhaps smoothing in this manner makes the diverse points harder to detect in the data.

The **Random** and **Grid** methods both do surprisingly well in terms of reconstruction error. The **Grid** method performing well is just random chance, dependent on the distribution of samples across the sample space. Perhaps the methods both happen to sample a point that is similar to large areas of the scene while not having a high entropy score. A point such as this would provide a large decrease in reconstruction error, even though the point itself is less interesting by some metrics.

The **Grid** method has no variance because, regardless of the noise added to the model, it chooses the same grid of points every time. As an alternative we could have changed the initialization point of the method, but we felt that this would have given us results nearly identical to the random method. As the true PIXL sampling scenario evaluates points in a sample in the same order each time, beginning at the top left of the sample set, we felt that it was appropriate to initialize this method in the same point as a baseline for comparison.

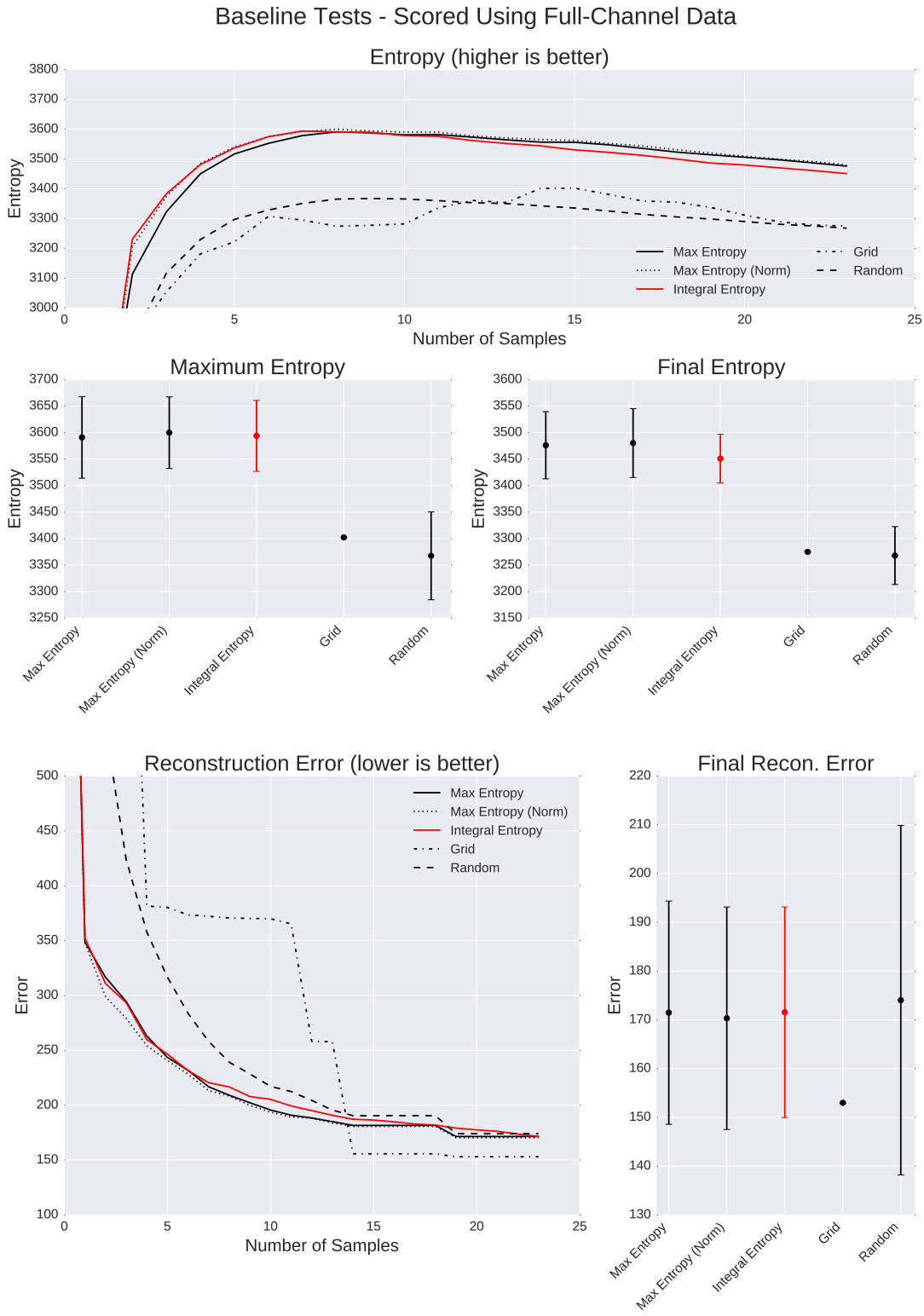


Figure 5.5: Entropy and reconstruction error for our baseline methods, scored on the full-channel data.

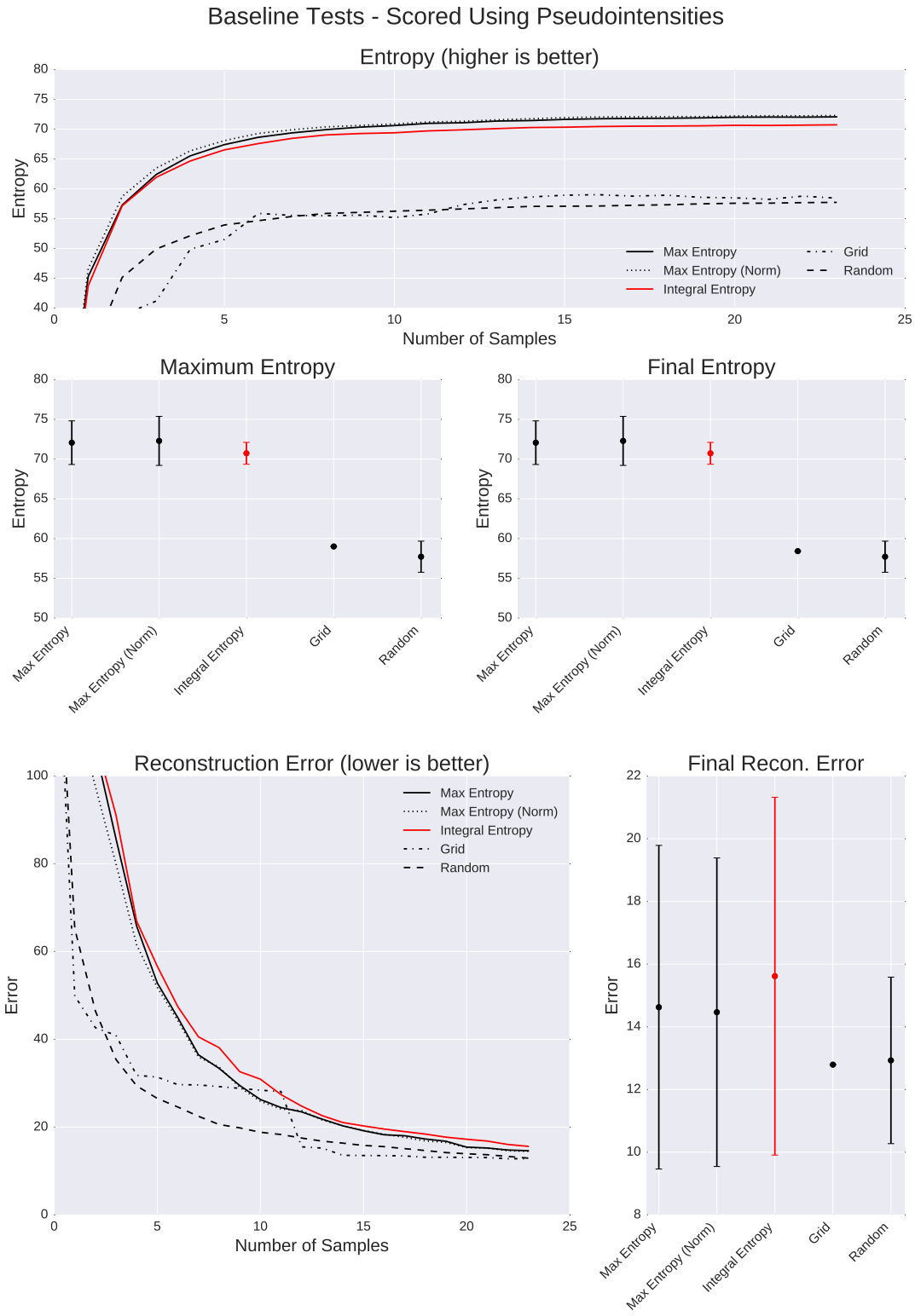


Figure 5.6: Entropy and reconstruction error for our baseline methods, scored on the reduced-channel dataset.

5.3.2 Adaptive Gaussian Mixture Model Results

The AGMM methods all perform competitively compared against the baseline approaches. Both **AGMM** and **AGMM-AddClasses** outperform all baseline approaches and **AGMM-P(Wrong)** in both entropy and reconstruction error, on both the full- and reduced-channel data, shown in Figures 5.7 and 5.8, respectively.

AGMM-AddClasses demonstrates that there is a strict improvement over the **AGMM** approach when classes are added dynamically using a Dirichlet process prior. Keeping the number of classes flexible results in both an increase in entropy and a reduction in reconstruction error.

The **AGMM** method gives similar, yet slightly inferior results to **AGMM-AddClasses**, which is somewhat expected as their selection methods are equivalent and only the number of classes change.

The **P(Wrong)** approach performs only adequately here under the entropy metric, being equivalent or worse than the greedy entropy maximization approach. However, it performs well when evaluated under the reconstruction error metric. There is no immediate intuition as to why there are these results, although perhaps it is because selecting points that do not have maximally high entropy scores means the algorithm is more likely to select points that are more similar to other points in the feature space, improving the reconstruction error score but scoring suboptimally with regards to entropy.

In all cases the simple model parameters of the AGMM perform surprisingly well. These parameters, such as a single common diagonal covariance matrix used by each mixture, simplify the models, yet they work well in this case where the points selected are specifically chosen to be in sparse regions of the feature space. If we instead had calculated the true covariance of the mixtures as multiple points are added, they would likely be constrained too heavily to make meaningful comparisons when merging mixtures and adding new points. In fact, in preliminary tests not shown here, we find that building a traditional GMM on sampled data runs into those exact problems, and performance suffers as a result.

It should be noted that the computational cost of computing the full density estimate of new points, as described in Equation 3.1, is very small given the limited fixed number of samples. Sillito et al. justify their use of the creation of Gaussian mixtures and their merging with the thought that it might be necessary to limit the number of "kernels" if the number of samples get too high. An obvious question, then, is how necessary is the creation and merging of mixtures as described in our approach, and what happens if we set the number of mixtures equal to the budgeted number of samples, effectively turning the AGMM method into KDE?

In fact, as shown in Figure 5.9, we have found that it is slightly advantageous to have a limited number of mixtures, as opposed to setting the number of mixtures equal to the number of samples. Perhaps this is because the addition of points to a mixture provides a better estimate of the covariance matrix so critical in the calculation of Mahalanobis distance, as mixtures with only a single point have a covariance matrix initialized with the σ^* estimate as described in Equation 3.3.

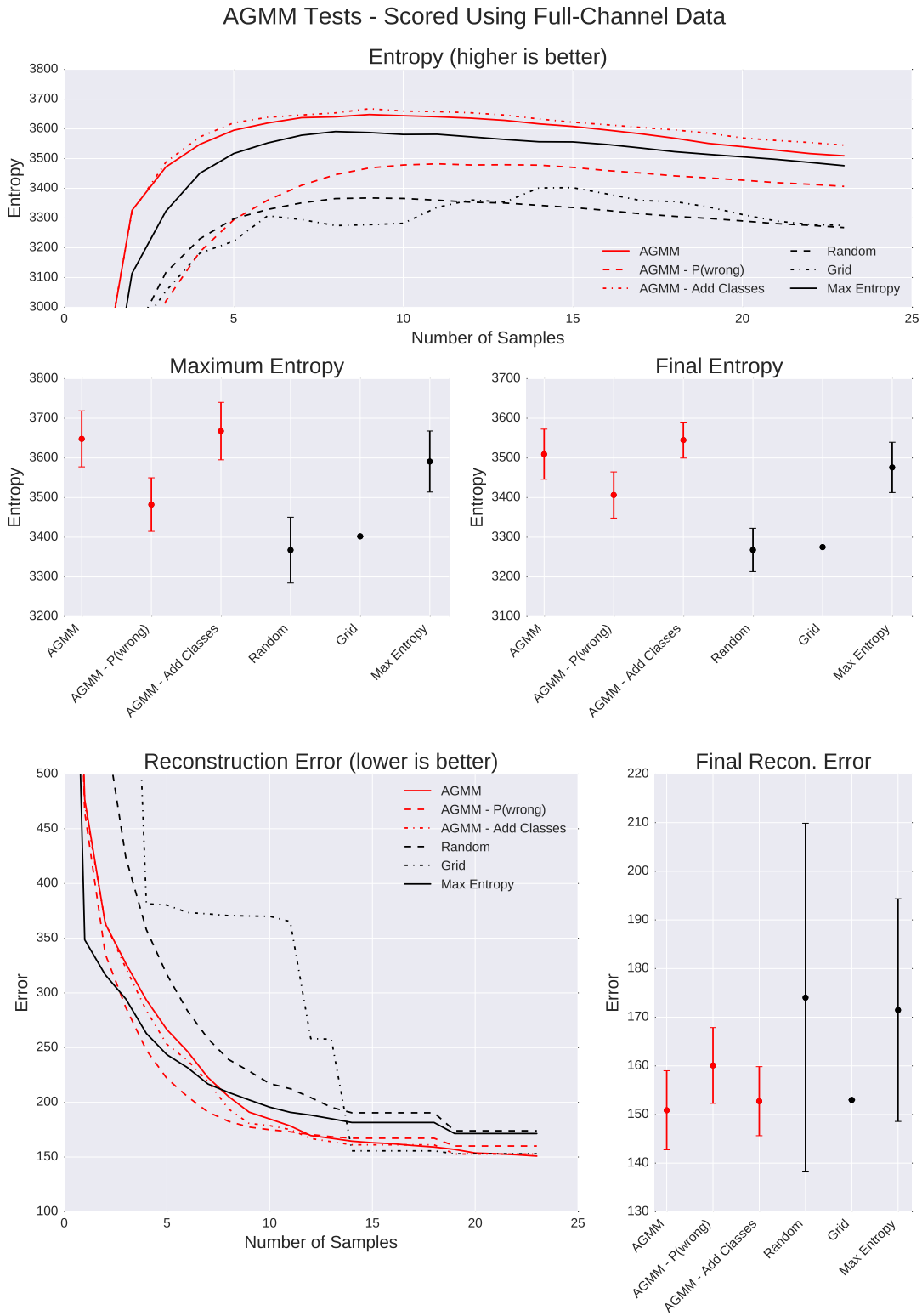


Figure 5.7: A comparison of the AGMM and baseline methods when scored against the full-channel data. The AGMM method that uses Dirichlet process priors to add new classes outperforms all others in both metrics.

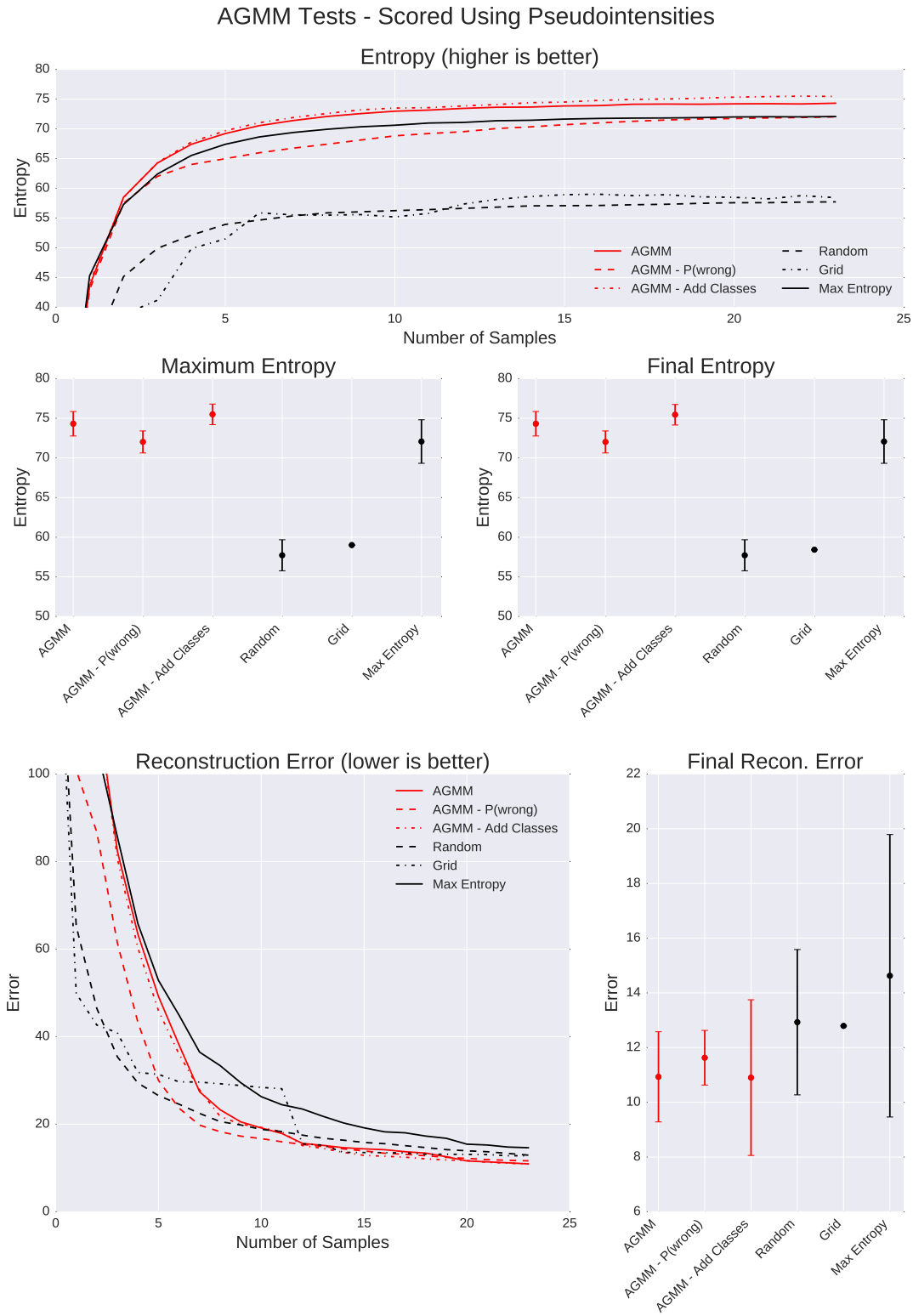


Figure 5.8: A comparison of the AGMM and baseline methods scores against the reduced-channel dataset that gives pseudointensities. The AGMM method that dynamically adds new mixtures outperforms both the base AGMM method and the method that selects points based on their probability of being wrong.

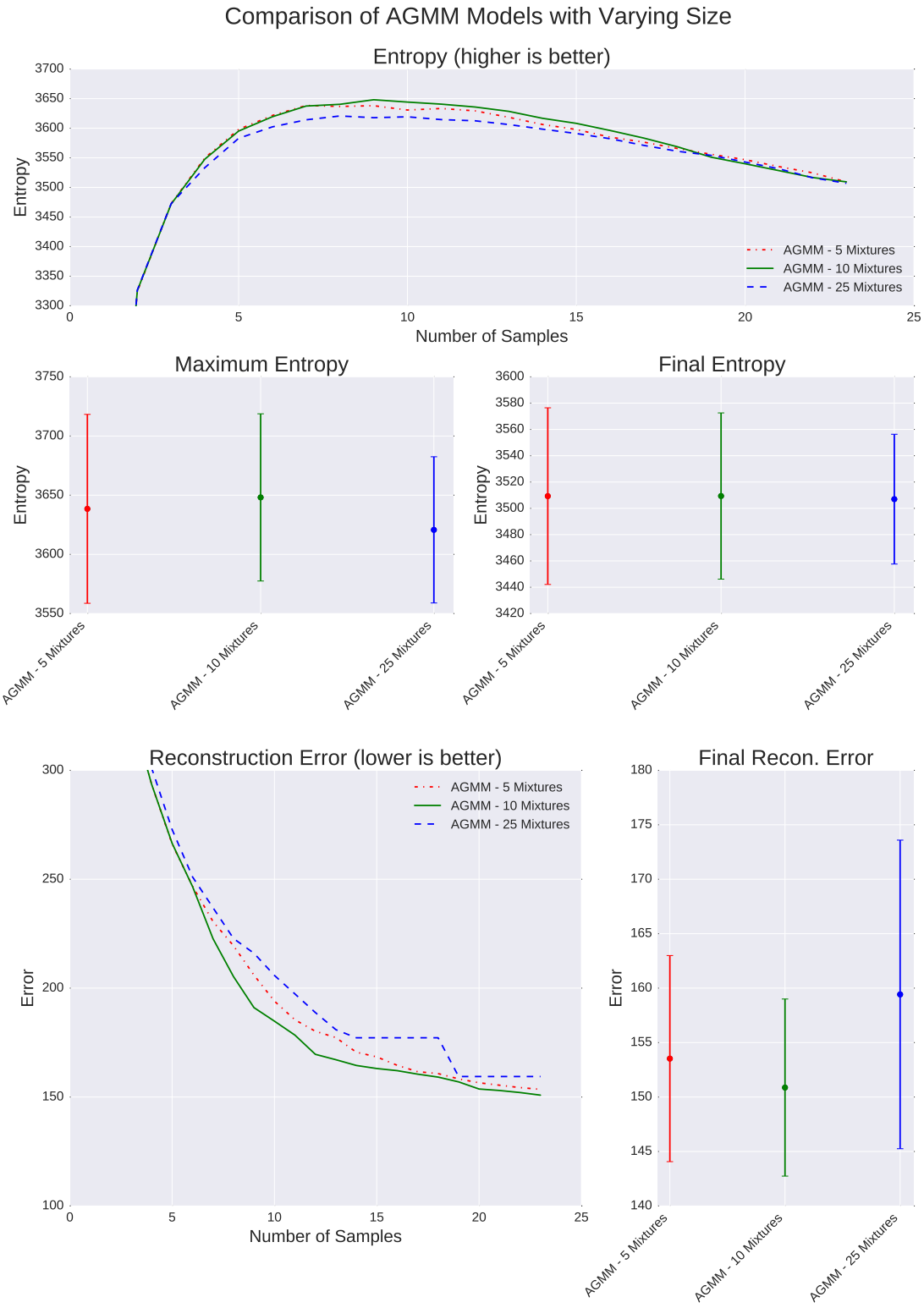


Figure 5.9: A comparison of the AGMM model using different numbers of clusters, tested with the full-channel data. There is a slight improvement from limiting the number of clusters, show by the 'AGMM - 10 Clusters' results, instead of always adding new points to new clusters, shown by the 'AGMM - 25 Clusters' approach.

5.3.3 Bayesian Optimization Results

All Bayesian optimization methods outperform the baseline methods with regards to the entropy scoring metric, indicating that they are well-equipped to find diverse points within the sample data. However, when compared to the baseline metrics in reconstruction error, only the **GP-MaxMean** method outperforms any of the baseline methods, and even then just barely. Results for the Bayesian optimization methods are shown in Figures 5.11 and 5.12.

The best performers out of the different acquisition functions are **GP-EI**, **GP-PI**, and **GP-MaxMean**. **GP-EI** achieves the highest overall entropy score out of all tested methods, AGMM methods included. However, both the **GP-EI** and **GP-PI** methods do poorly when considered under the reconstruction error metric.

In contrast, the **GP-MaxMean** method does well on all metrics. When scored using the entropy metric it outperforms the baseline on both the full- and reduced-channel data, indicating that using a Gaussian process for nothing other than a smoothing operation has some benefit. When scored using reconstruction error it outperforms all other methods. This is potentially because the smoothing operation leads it to sample fewer outlier points and instead sample points more representative of the overall data. Unlike the **Integral Entropy** method, the Gaussian process smoothing is perhaps better equipped to retain some of the interesting features present in the data.

The **GP-UCB** and **GP-MaxVar** methods are both somewhat unremarkable when compared to the other acquisition functions, especially considering their success in other domains. Both achieve mid-range entropy and reconstruction error scores on both the full- and reduced-channel data.

Finally, there is an interesting trade-off in the **GP-EI** and **GP-PI** methods when using the Matern or piecewise polynomial kernels, shown in Figure 5.10. In both of these methods, the Matern kernels reach their peak entropy scores faster than the polynomial kernels, around seven samples for the Matern kernel and ten for the polynomial kernels. While the final entropy scores of the Matern methods are worse, in cases where the number of samples is limited to fewer than seven, the Matern kernels offer the best performance.

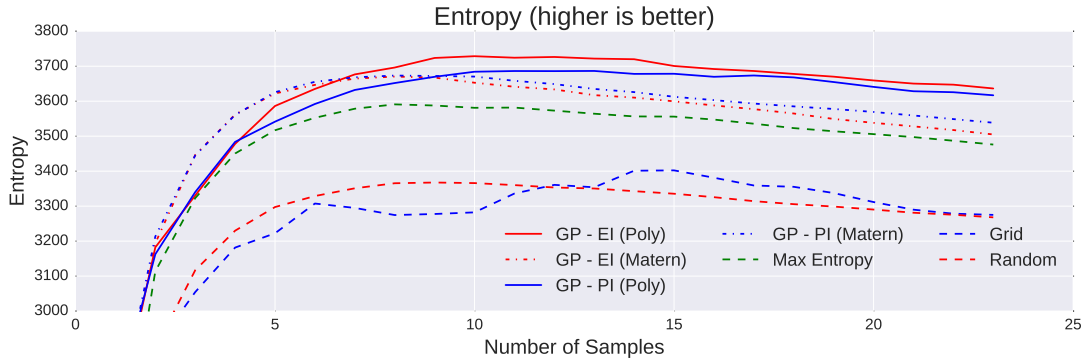


Figure 5.10: The Matern and piecewise polynomial kernels offer a trade-off between up-front and long-term entropy for the GP - EI and GP - PI methods.

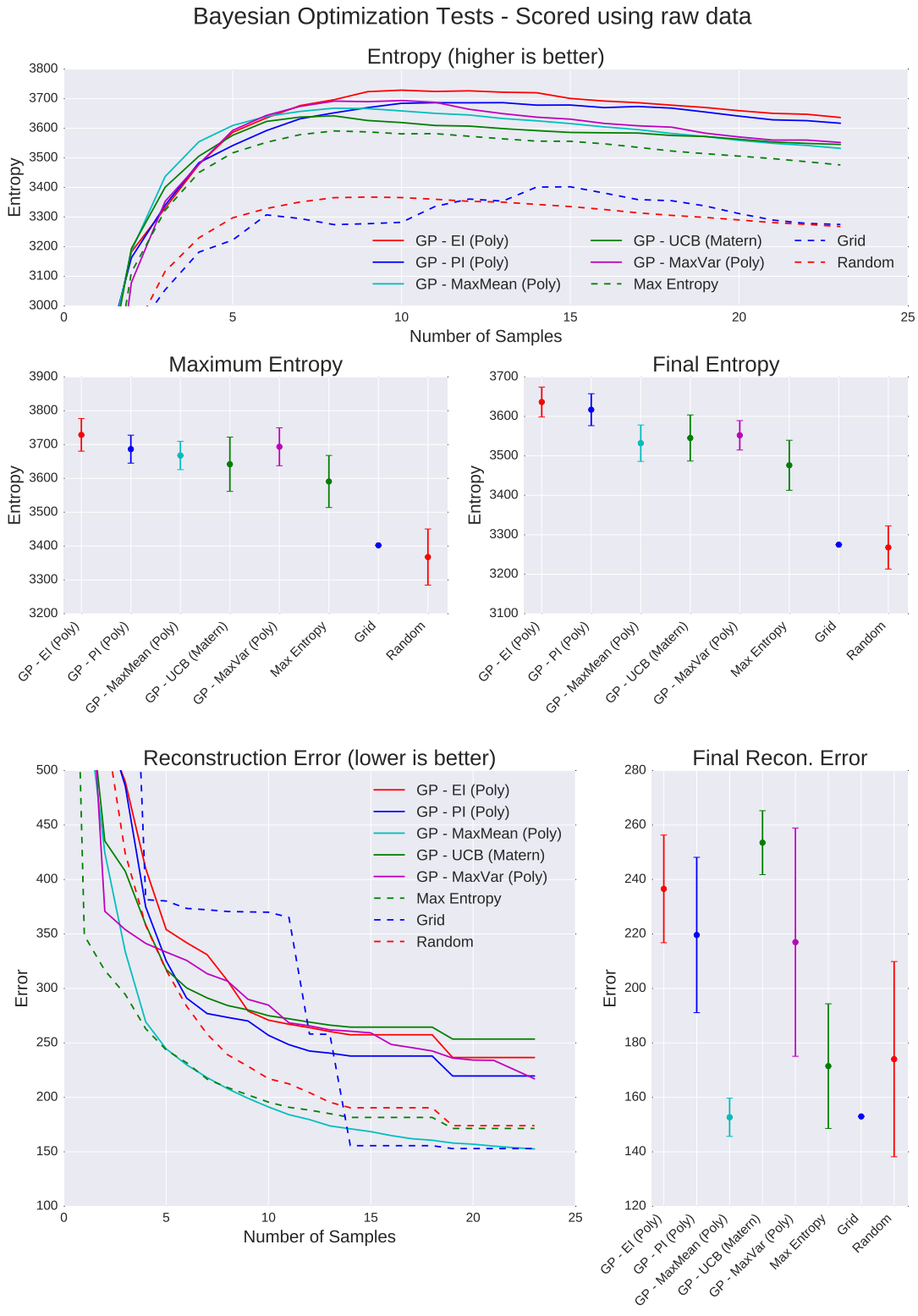


Figure 5.11: A comparison of the Bayesian optimization and baseline methods when scored against the full-channel data.

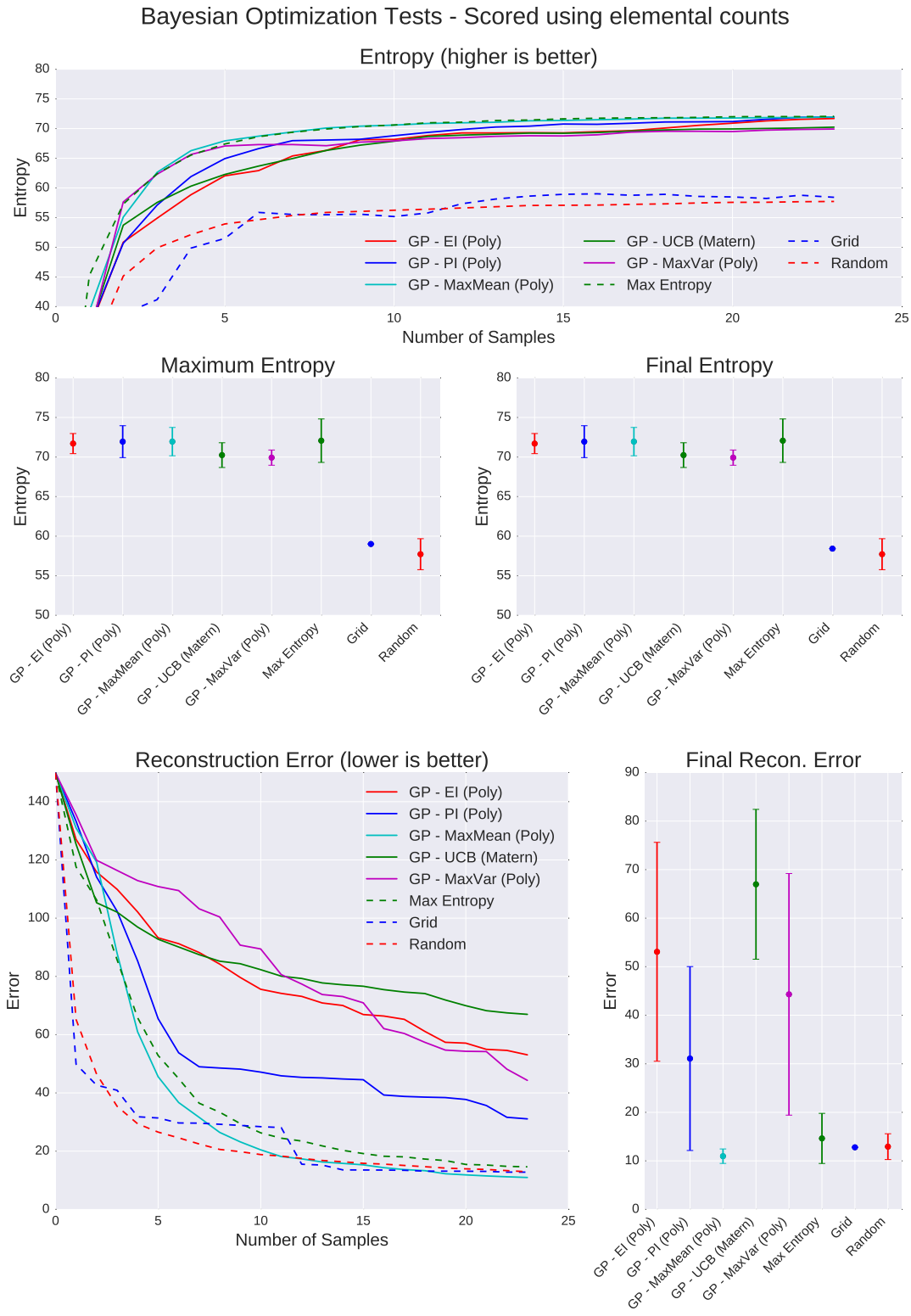


Figure 5.12: A comparison of the Bayesian optimization and baseline methods when scored against the reduced-channel data.

5.3.4 Discussion

Below are the overall results of our methods, scored on the full-channel data:

Method Name	Max Entropy	Final Entropy	Recon Error
Grid	3402.2	3274.9	152.9
Random	3367.5 ± 82.9	3267.9 ± 54.7	174 ± 35.8
Maximum Entropy	3590.9 ± 76.9	3476 ± 63.4	171.5 ± 22.9
Integral Entropy	3593.9 ± 67.0	3450.9 ± 45.7	171.5 ± 21.6
Probability of Improvement (GP-PI)	3686.5 ± 41.4	3616.9 ± 40.5	219.6 ± 28.5
Expectation of Improvement (GP-EI (Matern kernel))	3670.4 ± 45.0	3504.9 ± 59.9	198.4 ± 43.5
Expectation of Improvement (GP-EI (Polynomial kernel))	**3728.6 ± 48.1**	**3636.4 ± 37.9**	236.6 ± 19.8
Upper Confidence Bound (GP-UCB)	3641.9 ± 80.3	3545.3 ± 58.4	253.5 ± 11.7
Maximum Variance (GP-MaxVar)	3693.6 ± 56.1	3552.0 ± 37.2	217.0 ± 41.9
Maximum Mean (GP-MaxMean)	3667.8 ± 41.9	3532.1 ± 46.1	**152.7 ± 7.0**
Adaptive GMM (AGMM)	3620.7 ± 61.7	3506.9 ± 49.3	159.4 ± 14.2
AGMM - Add Classes (AGMM-AddClass)	3667.7 ± 72.4	3545 ± 45.3	**152.7 ± 7.1**
AGMM - P(Wrong) (AGMM-P(Wrong))	3482.3 ± 67.6	3406.5 ± 58.2	160.1 ± 7.8

Table 5.4: The overall results of our tests on the basic scenario. Our contributions and the best results are listed in bold.

We show the results of the best methods in Figures 5.13 and 5.14.

There is a common theme in many of our results, in that the methods that have the highest entropy scores often do not have as competitive of reconstruction error scores. As briefly discussed in the AGMM results discussion in Section 5.3.2, this likely occurs because there is a trade-off between sample diversity and sample coverage within our data. The most diverse samples in the dataset seem to be somewhat unique, in that sampling them provides less information about the rest of the sample space. However, the suboptimally-diverse points, or the points that are maybe 95% as good as the most diverse points, occur more often in the dataset or are more similar to large numbers of points in the dataset. Methods that sample these suboptimal points suffer slightly in their entropy scores, but see a large improvement in their reconstruction error score.

The exceptions to this theme, and the best performing methods, are the **AGMM**, **AGMM-AddClasses**, and **GP-MaxMean** methods. Each of these methods have high entropy scores and very competitive reconstruction error scores on both the full- and reduced-channel data.

Overall, these results lead us to focus our attention in the upcoming sections to the **GP-EI**, **GP-MaxMean**, and AGMM-based methods.

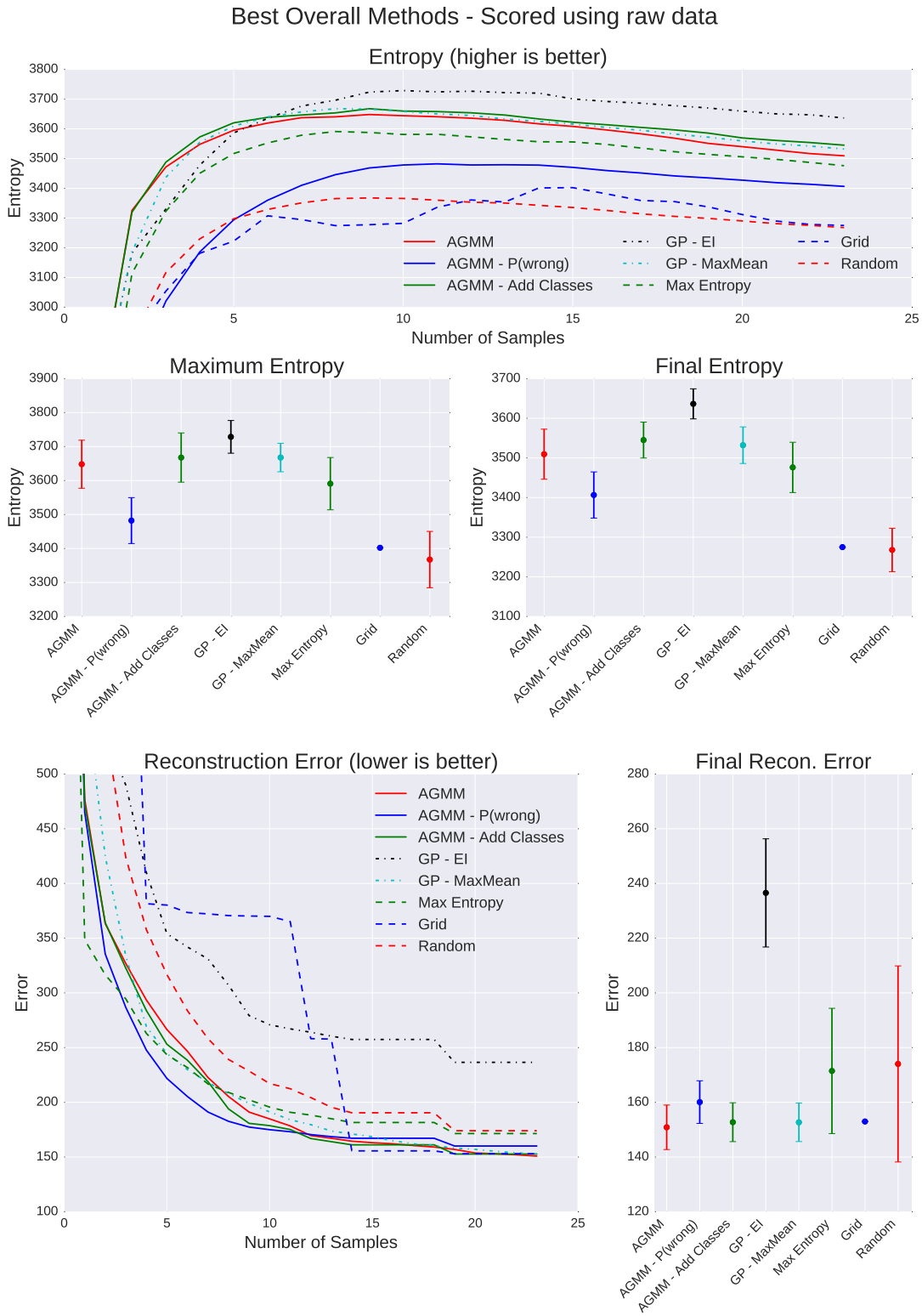


Figure 5.13: The best performing methods when scored against the full-channel data.

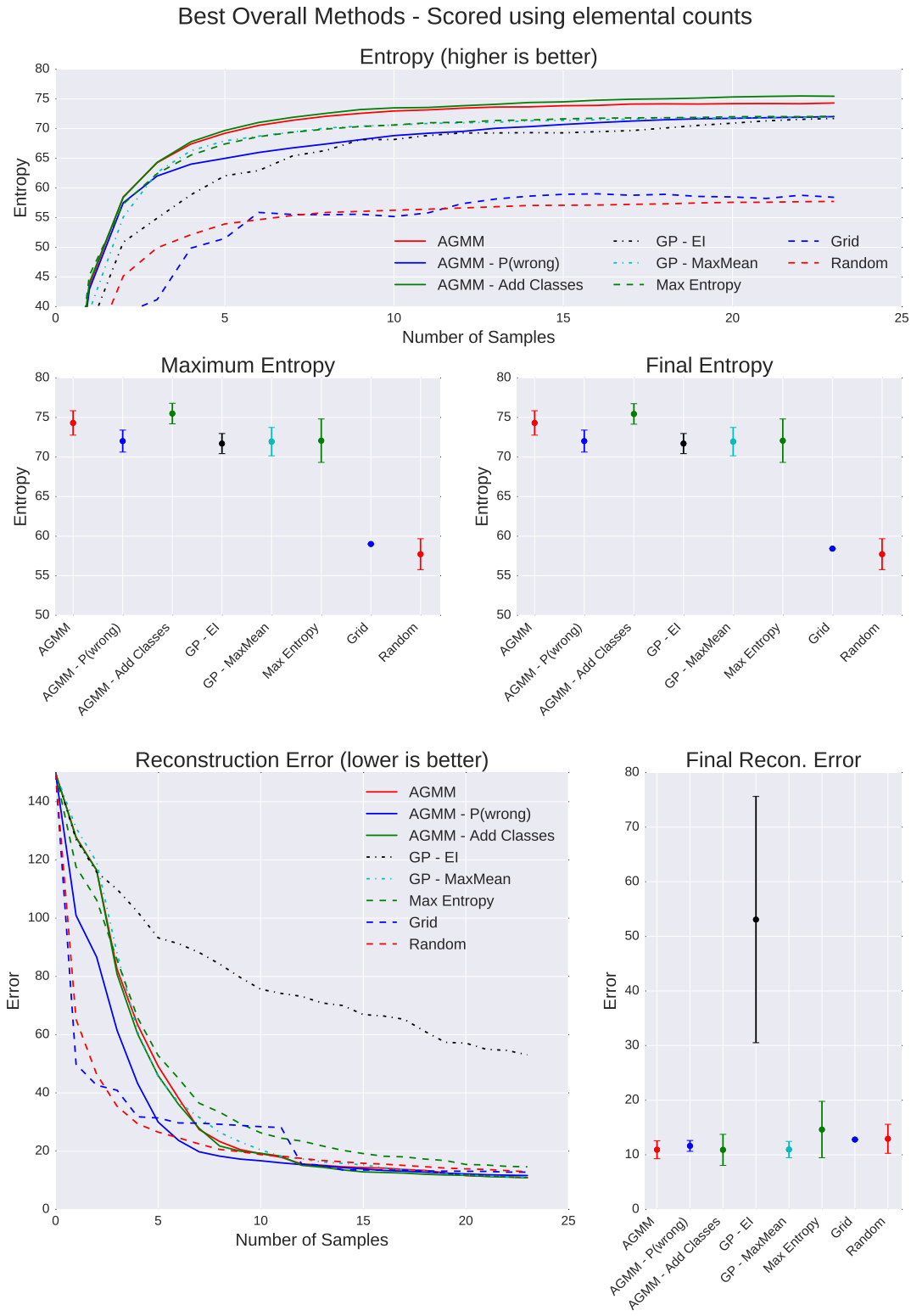


Figure 5.14: The best performing methods when scored against the reduced-channel data.

5.3.5 Detecting Rare Signatures

Approaches that achieve high entropy or low reconstruction error scores are good, but if these approaches are to be used in a planetary exploration mission then they must also be able to detect the rare signatures that are likely to offer value to a planetary geologist evaluating any sample data.

To evaluate each of the approaches under this metric, we first take the reduced-channel data and manually create abundance maps which mark regions in the data where each element in the reduced data is present in the scene. We consider an element to be present at a point in the scene if there is an observed value greater than or equal to 5.0. This value was chosen as it is likely that any point that has a value of 5.0 or greater actually has some amount of the given element at that location, and not just noise. If there was none of this element present at that location, then the expected Poisson statistics for the observation of that element have a mean of zero, making 5.0 statistically significantly far enough away that we can be confident there is some amount of the given element at this location.

Given the abundance maps, the rarity of each element is then calculated. Rarity is computed as the percentage of the image that has each element present. We then group elements into three rarity classes: Common (present in 10-100% of the image), Uncommon (1-10% of the image), and Rare ($< 1\%$ of the image). Examples of abundance maps are shown in 5.15.

Each approach's samples are then tested to see how often they sample Uncommon or Rare elements. Plots of the results of this are shown in Figure 5.16. The approaches are ordered from left to right based on their entropy scores in the previous section, in increasing order. The rare elements sampled by each approach are shown in Table 5.5.

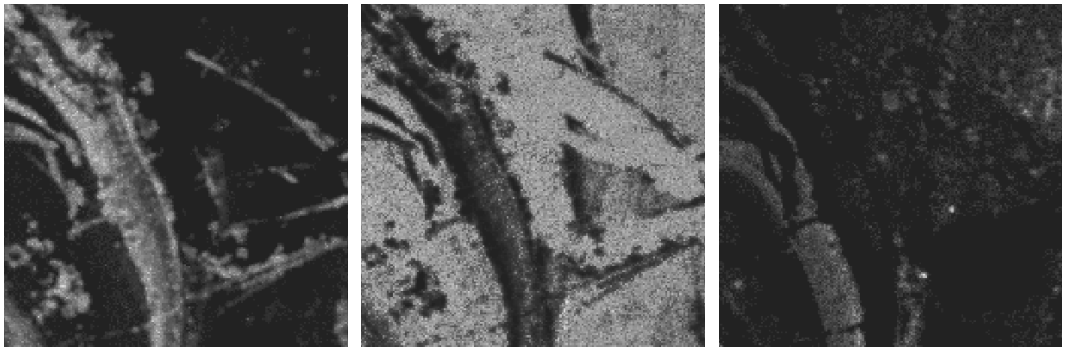
Method Name	Uncommon Elements	Rare Elements
Grid	P, Pr, Bi	
Random	K, P, Pr, C, Y, Bi, Be	
AGMM-P(Wrong)	K, P, Pr, C, Y, Bi, Be	Zr
Max Entropy	K, P, Pr, C, Y, Bi, Be	Zr, Cr
AGMM	K, P, Pr, C, Y, Bi, Be	Zr
GP-MaxMean	K, P, Pr, C, Y, Bi, Be	Zr, Cr
AGMM-AddClass	K, P, Pr, C, Y, Bi, Be	Zr
GP-EI	K, P, Pr, C, Y, Bi, Be	Cr

Table 5.5: Rare elements detected by each of the best approaches. Rare elements are those that appear in less than 1% of the test data.

The results show that all AGMM-based methods are good at finding rare signatures in the data. **GP-EI** samples, by far, the largest number of uncommon elements, but has a slightly lower count of rare elements sampled as compared to the other approaches.

It should also be noted that many of these approaches use entropy to score potential measurement locations. If the objective truly is to find the greatest number of rare elements or have the greatest number of samples that contain a rare or uncommon element, then perhaps a different

(a) Carbon (b) Praseodymium (c) Terbium



(d) Silver (e) Radium (f) Ruthenium

Figure 5.15: Abundance maps of different elements within the Dressermats PIXL dataset.

scoring metric should be used, or perhaps some of the more common elements should be ignored in the calculation of differential entropy.

As a side comment, many of the approaches here was also able to detect small quantities of Thulium and Palladium. These elements are not actually physically present in the sample, but are detected due to interference diffraction peaks from Barite and Sulfur. All other elements in the sample were confirmed to actually be present in the sample.

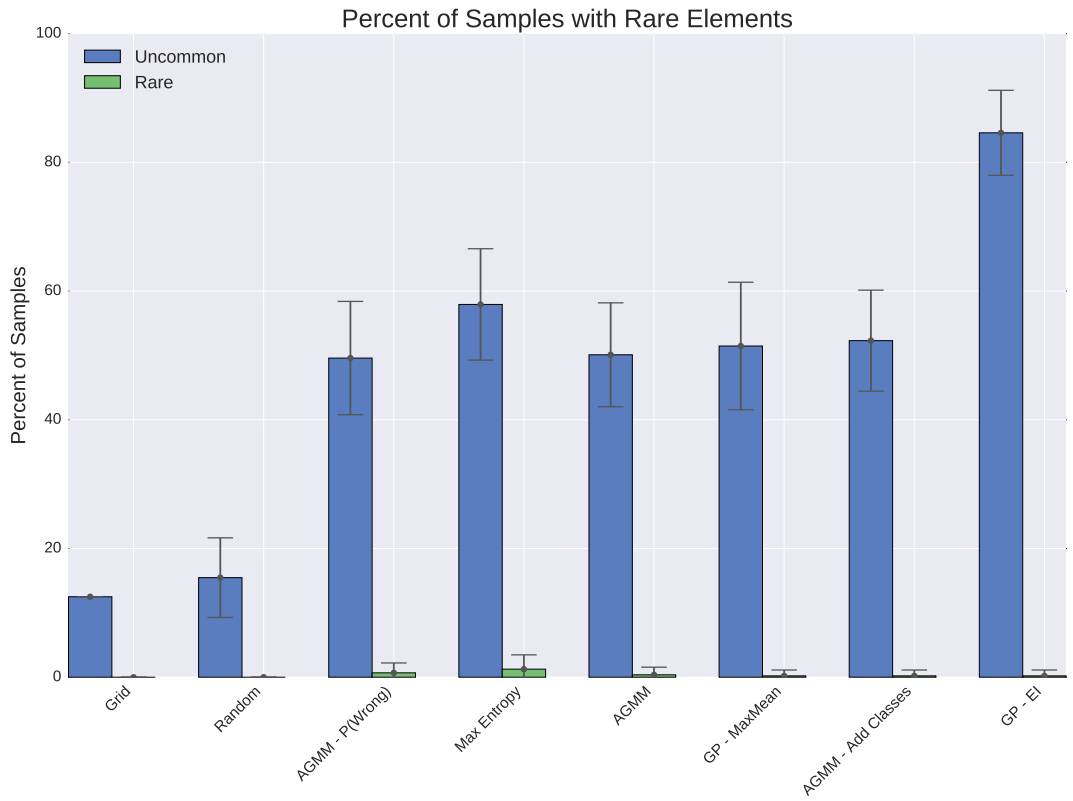


Figure 5.16: The percent of samples that contained Uncommon (blue) and Rare (green) elements, on average, by each approach. Approaches are ordered from left to right based on their entropy scores in the previous section, in increasing order.

Chapter 6

Adaptive Sampling with Limited Context

As Chapter 5 showed, adaptive sampling methods can improve how points are selected from contextual data. Both the AGMM- and Bayesian optimization-based approaches showed improvement over greedy and basic methods of sample selection.

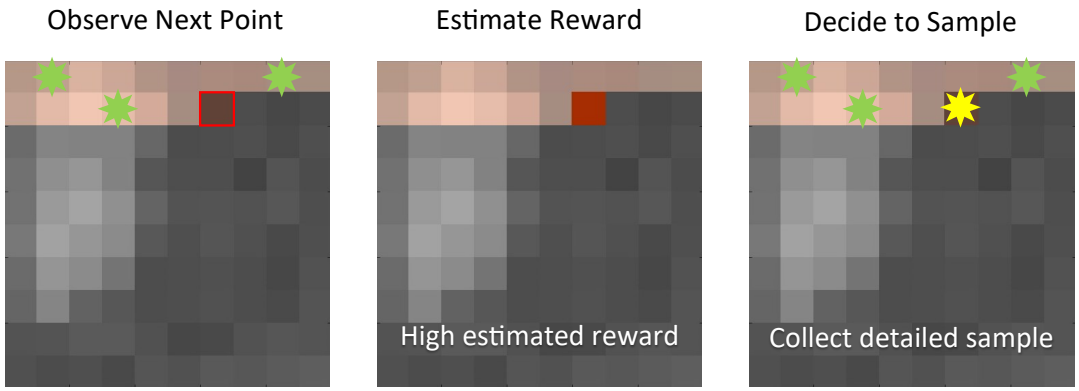
However, many robotic sampling situations do not have the full contextual information available beforehand, as it was for the experiments in Chapter 5. That data assumed that, for every potential sample point, we had noisy, yet approximate contextual data, and that the agent was able to freely sample any point with no constraints.

This scenario is unlikely for a robotic explorer because of time and resource constraints on going anywhere or everywhere. It is also unlikely with observers like orbital satellite or microscopic field imagers, because it is difficult or impossible to return to the exact location of a prior observation. If we consider a robot sampling a surface with some directed instrument, such as a targeted spectrometer, any number of issues, such as backlash within gears, vehicle motion, or limitations on aiming accuracies, lead to errors between the contextual data and the sampling data. Specifically, the rover can collect and observe the contextual data at any number of points, but there are no guarantees that the rover can return the instrument to that exact location to take a more detailed follow-up measurement.

The second experimental scenario considers how to handle sampling in the microscopic domain. In this scenario, we consider the case of a robot using a spectrometer on the microscopic scale, in which measurements are millimeters apart. Data is collected in a raster pattern, and we must evaluate the quickly-collected and noisy data at that point decide immediately whether or not to collect a detailed spectrum at that location [Thompson et al., 2014].

In contrast to the experiments in Chapter 5, in this section contextual data is limited to only what the rover has seen thus far in the raster. Furthermore, this contextual data is made up of noisy points, which the rover has not fully sampled, and a handful of more accurate points at which the rover elected to follow up with more detailed measurements. The contextual data can be used to

build up models in order to improve later selections, but once the agent has decided to sample or not to sample at a location, there is no returning. An example of the process is shown in Figure 6.1.



(a) Choosing the third sample point

(b) Choosing the fourth sample point

Figure 6.1: An example of the sampling process in the microscopic case. Left column: Only the next point in the scan is observed. Center: A rewards is estimated given prior samples. Right: The rover chooses to sample the point or pass.

This scenario is based heavily on the operational paradigm of the Planetary Instrument for X-Ray Lithochemistry (PIXL). As discussed in 5.1, PIXL is an X-ray fluorescence spectrometer that collects hyperspectral measurements at the millimeter scale. It is designed to operate on the end of the Mars 2020 rover's arm. Because of the mobility limitations of the arm and the PIXL instrument itself, it is difficult to return exactly to a location that was already sampled. Thus, the instrument must be equipped with adaptive sampling techniques that make it able to discern, in real time, whether or not to further integrate over a location to reduce the noise at interesting points.

This chapter details our approach to this microscopic sampling scenario, and the applicability of and changes to the methods outlined in Chapter 5. We evaluate the approaches on data collected

by the PIXL instrument, and this experiment mimics the same operational strategies as might be encountered on a use of the PIXL instrument on Mars.

6.1 Microscopic Sampling with the PIXL Instrument

The PIXL instrument is an X-ray fluorescence spectrometer that collects a raster of spectra at the sub-millimeter scale, as discussed in Section 5.1. The spectrum collected contain 4096 energy channels that range from 0-41 KeV. The instrument can collect spectrum spectrum quickly, obtaining a noisier measurement of the surface, or can integrate more fully and increase the signal to noise ratio (SNR) at that location. A quick scan takes five to ten seconds, but a full integration can take minutes [Thompson et al., 2015a].

A primary concern of the instrument is how to collect the most interesting and informative samples with limited time. A full raster might be over tens of thousands of measurement locations, so it quickly becomes infeasible to fully sample every location. Instead, a small subset must be selected for detailed sampling, the size of which is relative to a pre-determined budget.

The experiments in this section are evaluated using the same constraints as those the instrument will face on the Mars 2020 mission. We consider a fixed budget that limits the total number of long-dwell spectra to a budget B: $|X_{sampled}| < B$. The overall performance of all methods are then compared over the diversity and coverage of the fully sampled points, as calculated using the entropy and reconstruction error metrics discussed in Section 3.3.

This scenario is similar to that of stream-based sampling in the active learning field [Settles, 2010]. In pool-based active learning, the full set of points is available for analysis prior to selection, similar to the experiments in Chapter 5. In stream-based active learning, samples arrive one at a time and the agent must decide whether or not to query the oracle for the true class of the point.

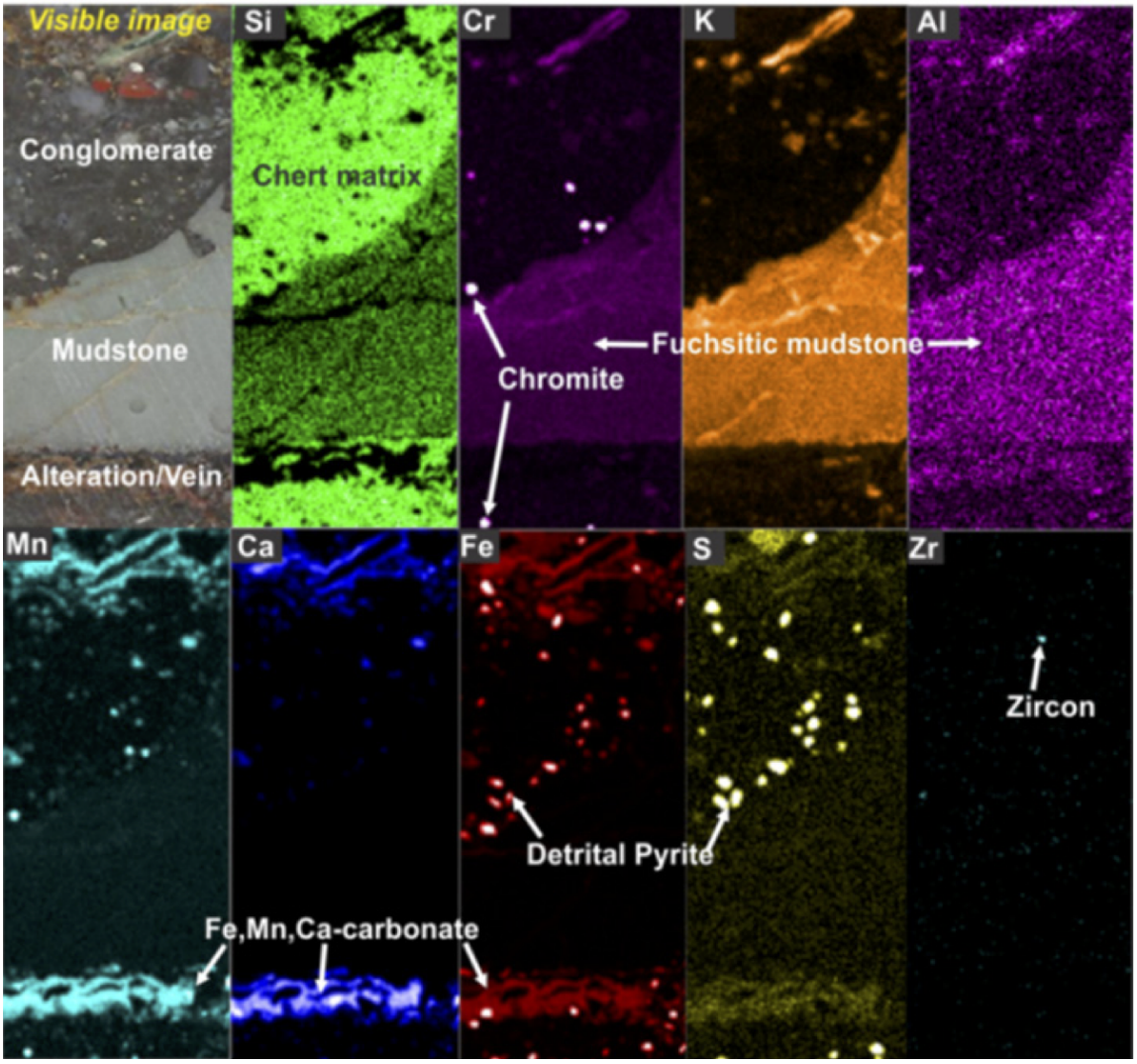
We again test approaches on the dataset described in Section 5.1. As in that section, data is post-processed to convert it from X-ray energies to pseudoabundances representing the abundance of specific elements, using the method discussed in Section 5.1.1. A visual example of a similarly post-processed dataset is shown in Figure 6.2.

6.1.1 Prior Work in Adaptive Selection for the PIXL Instrument

There are two main works that have dealt with this exact scenario using PIXL data. In the first, Thompson et al. describe a method for using vector quantization theory for selecting a representative subset of the data [Thompson et al., 2014]. Thompson frames this problem as the selection of M long dwell points, $B = \{b_1, ..., b_M\}$, within a transect $A = \{a_1, ..., a_n\}, B \in A$,. They define a distance measure of a new point to the existing samples as the cost:

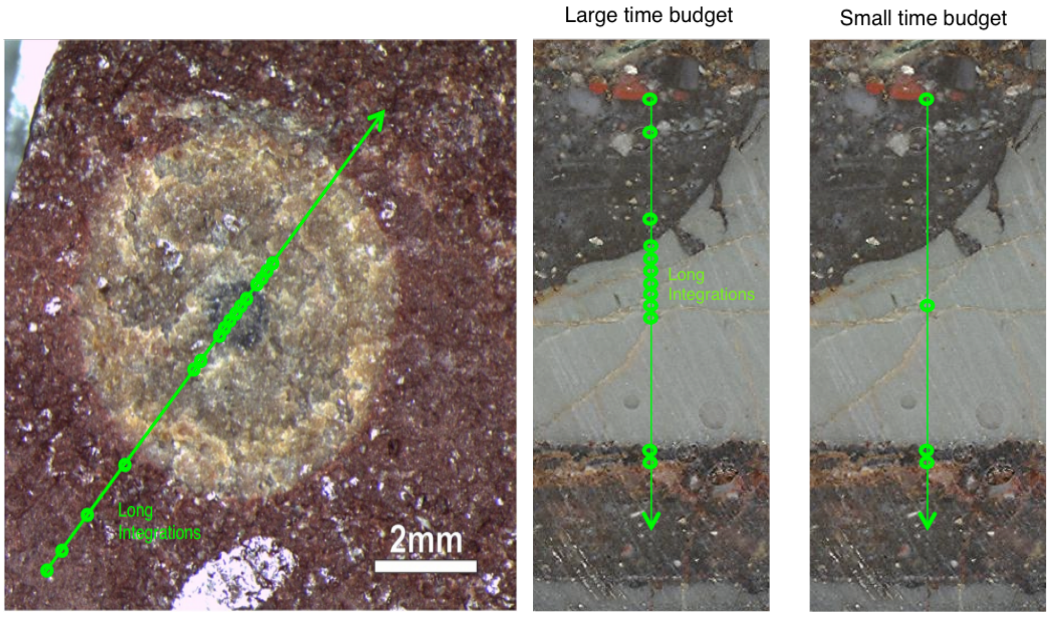
$$L(B) = \sup_{a}(\min_{x \in B} |w^T a - w^T x|_2) + C(B) \tag{6.1}$$

in which w is a weight vector used to emphasize or de-emphasize specific elements in the PIXL data, and $C(B)$ is a cost function that increases monotonically with the number of fully sampled points.



(a)

Figure 6.2: An example of the different elemental abundances present in a sample, as measured by the PIXL instrument. The image in the top left shows a color image of the sample. All other images show the same scene with listed element colorized to show its relative abundance. From [NASA, 2016e].



(a) (b)

Figure 6.3: Examples of long dwell point selected on transects by Thompson et al.'s vector quantization approach [Thompson et al., 2014].

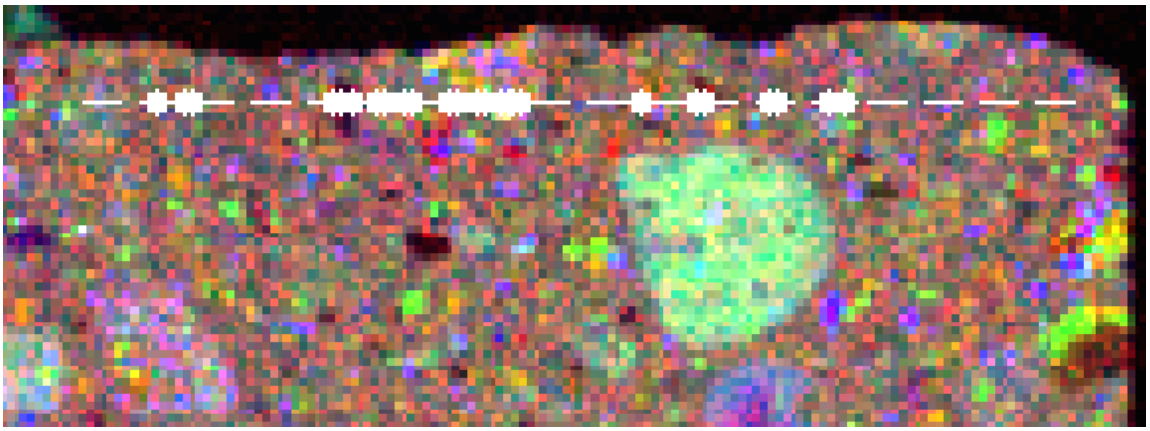
As new data is collected, they decide to sample if the weighted distance from a new sample point, a, from the existing long-dwell points is greater than a threshold T:

$$\text{trigger if } \min_{x \in B} |w^T a - w^T x|_2 \text{ and } |B| < M \tag{6.2}$$

This approach has the advantage that there are a number of tunable parameters available to the scientists operating the instrument. First, they can adjust the weights on the elements in order to change the priority of some elements over others. Second, they can change the threshold T, to adjust the rate of data collection. A lower value of T will naturally lead to more samples collected, which may be appropriate depending on the time allocated for a sample. Finally, the scientists can adjust the budget, M, to make sure the allocated number of samples is not exceeded.

An example of their approach is shown in Figure 6.3. All images assume a direct transect across the sample, with long-dwell points selected along the way. This is slightly different from the raster approach we outlined above, in which the instrument collects a number of linear transects across a sample in a grid-like pattern, unlike here where only one transect is collected, but the approach can easily be extended to work in the raster setting as well.

The second main work on adaptive sampling methods for PIXL data is that of Thompson et al. [Thompson et al., 2015a]. They consider both adaptive sampling methods and visualization methods for interpretation of the data. In their adaptive sampling work, they consider the same



(a)

Figure 6.4: An example of long dwell points selected on a transect by Thompson et al.'s interval-based approach [Thompson et al., 2015a].

scenario we discuss here, in which the rover must view a quick scan of a point and decide whether or not to use a budgeted number of follow up operations to gather more information at that point.

They propose two main methods for adaptive selection. First they discuss a density-based approach based on their prior work in [Thompson et al., 2014]. Here, they maintain a library of the long-dwell spectra collected prior, v, then fully sample a new point, u, if its distance to the nearest point in the library exceeds a threshold, τ:

$$\text{trigger when } \exists\ v \text{ S.T. } d(v,u) > \tau \tag{6.3}$$

$$d(v,u) = \sqrt{\sum_{i=0}^{N} \left(\frac{u_i - v_i}{w_i}\right)^2} \tag{6.4}$$

Similarly to what they propose in [Thompson et al., 2014], each elemental channel has a weight w_i, that can be adjusted a priori by the scientists operating the instrument, allowing them to focus on elemental signatures of more interest to the science team.

The second method they propose is an interval-based approach. Here, they consider triggering a long-dwell spectrum if the value of any element in the quick sample exceeds a threshold. This threshold is defined as:

$$\text{trigger when } \exists\ i \text{ S.T. } u_i \notin Q \tag{6.5}$$

Q can be any interval, but they choose to define it in terms of standard deviations from the mean spectrum collected thus far. In their work they use the interval $\Phi_i = [-4, 4]$, meaning they sample any point in which a single channel is more than four standard deviations away from the mean spectrum. They also indicate that the intervals can be used to match signatures, for example sampling points that are less than one standard deviation away from a signal known to be interesting.

Their work finds that the interval-based approach produces the greatest entropy of the resulting sample set. An example of the long-dwell spectra collected on one of their transects is shown in Figure 6.4.

6.2 Adaptive Methods for Stream-Based Sampling

In the results of the prior chapter where each approach had access to the full contextual data, we demonstrated a number of different adaptive sampling methods that improve sample diversity and coverage, as compared to the baseline metrics. Unfortunately, only a handful of these methods extend well to this scenario.

The primary reason for this is that the Bayesian optimization approaches, such as **GP-EI**, require a number of training points in order to reconstruct an accurate estimate of the reward at candidate sampling points. In this scenario, in which future data is unobserved, and prior data is initially limited, Gaussian process methods will perform poorly until there is enough training data to build a representative model. While that might be acceptable if a sample has tens of thousands of points and the first few hundred can be ignored to build up training data for the Gaussian process, we do not want to limit ourselves to that case and ignore cases where smaller transects are used to quickly characterize a surface.

The AGMM methods have difficulties with this scenario as well. The AGMM model is build with a fixed, pre-selected number of mixtures, K. It then always fully samples the first K points in order to build up these mixtures. In this scenario, K might represent a prohibitively large amount of the budget, sampling suboptimal points just to build up the model. Even in the **AGMM-AddClasses** case, where classes are dynamically added using a Dirichlet process prior, the probability of adding a new class is equal to the probability of being wrong, effectively reducing the method in this case to be equal to the **AGMM-P(Wrong)** approach.

As a result, we only consider the **AGMM-P(Wrong)** approach in this scenario. We find this to be an intuitive approach to sampling, and can almost be thought of as an extension to the interval-based approach outlined above. We provide more details on the implementation below.

6.2.1 Misclassification Estimation Using Dirichlet Priors

In the base AGMM implementation, the model is formulated as a simplified Gaussian mixture model (GMM), in which there are a fixed number of mixtures that are built up over time.

This formulation lends itself well to finding diverse points, as shown in Chapter 5 where approaches were given access to the full contextual data. However, the basic formulation is a poor match for this scenario, in that the first K points are always sampled, where K is the count of mixtures pre-allocated in the AGMM model. This is potentially a waste of the first K samples of the budget.

Instead, we consider the AGMM - P(Wrong) formulation. In this model, we still have a fixed number of K mixtures. However, we are a little more restrictive with the points that we sample,

and instead favor only sampling points that are dissimilar from the data we have already collected, similar to the interval-based approach above.

Specifically, we sample points based on $P(wrong|x)$, defined as

$$P(wrong|x) = 1 - P_n(c'|x) \tag{6.6}$$

$$c' = \arg\max_{c \in C} P_c(x|c) \tag{6.7}$$

where $c \in C$ are the existing mixtures in the model, and P_n is the Dirichlet-weighted probability of belonging to an existing class or a new, unobserved class:

$$P_n(c \in \{C \cup new\}|x) \propto \begin{cases} \frac{m_c}{n-1+\alpha} P_c(x|c) & \text{if } c \in C \\ \frac{\alpha}{n-1+\alpha} P(x) & \text{if } c = \text{new} \end{cases} \tag{6.8}$$

At every candidate sample point, we calculate the value of $P(wrong|x)$, then decide whether or not to sample at that point. For all tests the sampling criterion is

$$\text{trigger when } P(wrong|x) > \tau \tag{6.9}$$

In this experiment we choose $\tau = 0.997$, chosen to be similar to the interval-based approach with an interval of $[-3\sigma, 3\sigma]$. Finally, when points are found to be informative enough, they are fully sampled and they are added to a completely new mixture.

We further take advantage of the AGMM structure by merging in the noisy contextual data, even if the point is not sampled. This data is forced to merge in with existing clusters, even if the number of points sampled is below K, the maximum number of mixtures in the AGMM model. This has the advantage of building up more representative mixtures over time and improves results when compared to not using contextual data.

6.2.2 Baseline Methods

To compare against **AGMM-P(Wrong)** approach, three baseline methods are considered.

The first is the interval-based approach by Thompson et al. [Thompson et al., 2015a], defined as:

$$\text{trigger when } \exists\, i \text{ S.T. } u_i \notin Q_i \tag{6.10}$$

Q can be any interval, but they choose to define it in terms of standard deviations from the mean spectrum collected thus far. In their work they use the interval $\Phi_i = [-4, 4]$, and we test with both $\Phi_i = [-4, 4]$ and $\Phi_i = [-3, 3]$. We refer to these methods as **Sigma4** and **Sigma3**, respectively. The mean spectrum used in the tests is composed of all points sampled thus far, both noisy and long-dwell spectra.

The next approach is a random sampling approach, in which the long-dwell locations are chosen randomly throughout the set of samples. This approach is referred to as **Random**.

Finally, we compare against a periodic sampling approach. This approach chooses long dwell spectrum at a periodic interval spanning the sample set, the length of which is relative to the budgeted number of samples allowed. This approach is referred to as **Grid**.

6.2.3 Testing Methodology

We test all approaches in a similar manner to the methodology used for the general case in Section 5.2.4.

Specifically, we generate a number of noisy datasets, based on the Poisson statistics of the original PIXL data. Because the data is Poisson distributed, we can assume that the mean and variance of each channel at each point is equal to the value of the channel point itself:

$$\lambda_{i,b} = \text{PIXL}(x_i, b) \tag{6.11}$$

$$\tag{6.12}$$

in which b is the channel that describes the elemental response at this point, and x_i is a location within the image. To create a noisy version of the data at this location and for this element, we draw a sample from a Poisson distribution using the expected value, $\lambda_{i,b}$:

$$\text{Noisy}(x_i, b) \sim \text{Poisson}(\lambda_{i,b}) \tag{6.13}$$

In this manner we create 100 noisy approximations of the original dataset. These noisy approximations are then transformed to elemental counts using the approach outlined in Section 5.1.1.

All tests are initialized by sampling at the top left point in the data, with a total sampling budget of 20 points. Upon choosing to sample a point, the full data is observed. The scoring metrics, described in Section 3.3, are based on the values of the long dwell data in this set of clean samples. The final results are averaged across all tests.

6.3 Results and Discussion

Random and **Grid** do very well in this experiment, especially in their initial scores, where the number of samples is five or fewer. This is likely because both the random and periodic approaches cover a larger space of the data, as they are designed to select samples that can potentially span the entire raster. In contrast, the other methods have somewhat of a slower start, as they must build up initial models over the section of the raster where the initialization point is located, so their long-term results are higher, but their initial scores are lower.

This is especially true for the interval-based approaches, **3Sigma** and **4Sigma**. These approaches rely on the prior points sampled to gradually build up a model, which we can directly observe in the results. In fact, while all other methods have plateaued within 15 samples, it appears that the interval-based approaches are still improving even when the budget has been exhausted. Over time, they become more and more selective about which points to select for long-dwell samples, but the initial points selected are relatively poor. The **AGMM-P(Wrong)** approach also builds up a model over time, but does a much better job of selecting initial points.

From these results we see that the overall best performer is the **AGMM-P(Wrong)** method. It achieves the best scores in all but two metrics, and even then is close to the optimal score.

The full results of the baseline and **AGMM-P(Wrong)** methods are shown in Figures 6.6 and 6.7.

We also evaluate the number of points each method considered before exhausting its sample budget, shown in Figure 6.5. We ignore the **Random** and **Grid** methods, as they are designed to sample points across the entire sample space. We see that the **AGMM-P(Wrong)** case, on average, considers more points for evaluation than either the **3Sigma** or **4Sigma** methods. This, in an of itself, is interesting, as it means that the **AGMM-P(Wrong)** method is able to build up a better model of the data than the alternative approaches. This is perhaps due to the sampling threshold selected, $P(Wrong) > 0.997$, and likely partially from the fact that the unsampled data is used to improve the model in a more effective manner than in the interval-based approaches.

Below are the tabled results of experiments, scored on the full- and reduced-channel data:

Method Name	**Max Entropy**	**Final Entropy**	**Recon Error**
Grid	3347.4 ± 0.0	3278.4 ± 0.0	204.5 ± 0.0
Random	$\mathbf{3357.0 \pm 43.5}$	3286.8 ± 34.6	202.9 ± 23.5
Interval (3 Sigma)	3253.2 ± 47.1	3253.2 ± 47.1	174.6 ± 4.0
Interval (4 Sigma)	3252.4 ± 34.4	3252.4 ± 34.4	180.6 ± 14.4
AGMM - P(Wrong) (AGMM - P(Wrong))	3352.3 ± 64.0	$\mathbf{3332.8 \pm 50.7}$	$\mathbf{162.7 \pm 3.9}$

Table 6.1: The overall results of tests on the microscopic sampling scenario, scored against the full-channel data. Our contributions and the best results are listed in bold.

Method Name	**Max Entropy**	**Final Entropy**	**Recon Error**
Grid	57.5 ± 0.0	57.5 ± 0.0	13.5 ± 0.0
Random	57.7 ± 2.6	57.7 ± 2.6	15.3 ± 4.4
Interval (3 Sigma)	61.7 ± 1.3	61.7 ± 1.3	$\mathbf{11.8 \pm 0.6}$
Interval (4 Sigma)	61.3 ± 3.3	61.3 ± 3.3	12.7 ± 1.1
AGMM - P(wrong) (AGMM - P(Wrong))	$\mathbf{66.1 \pm 1.4}$	$\mathbf{66.1 \pm 1.3}$	11.9 ± 1.4

Table 6.2: The overall results of tests on the microscopic sampling scenario, scored against the reduced-channel data that produces pseudointensities. Our contributions and the best results are listed in bold.

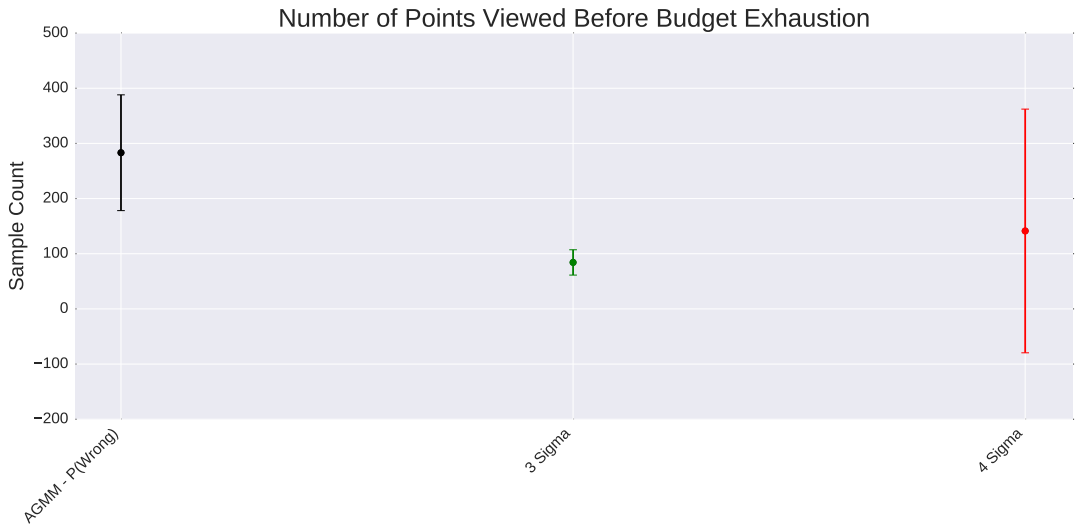


Figure 6.5: Count of the number of samples each method typically evaluated before exhausting its sample budget.

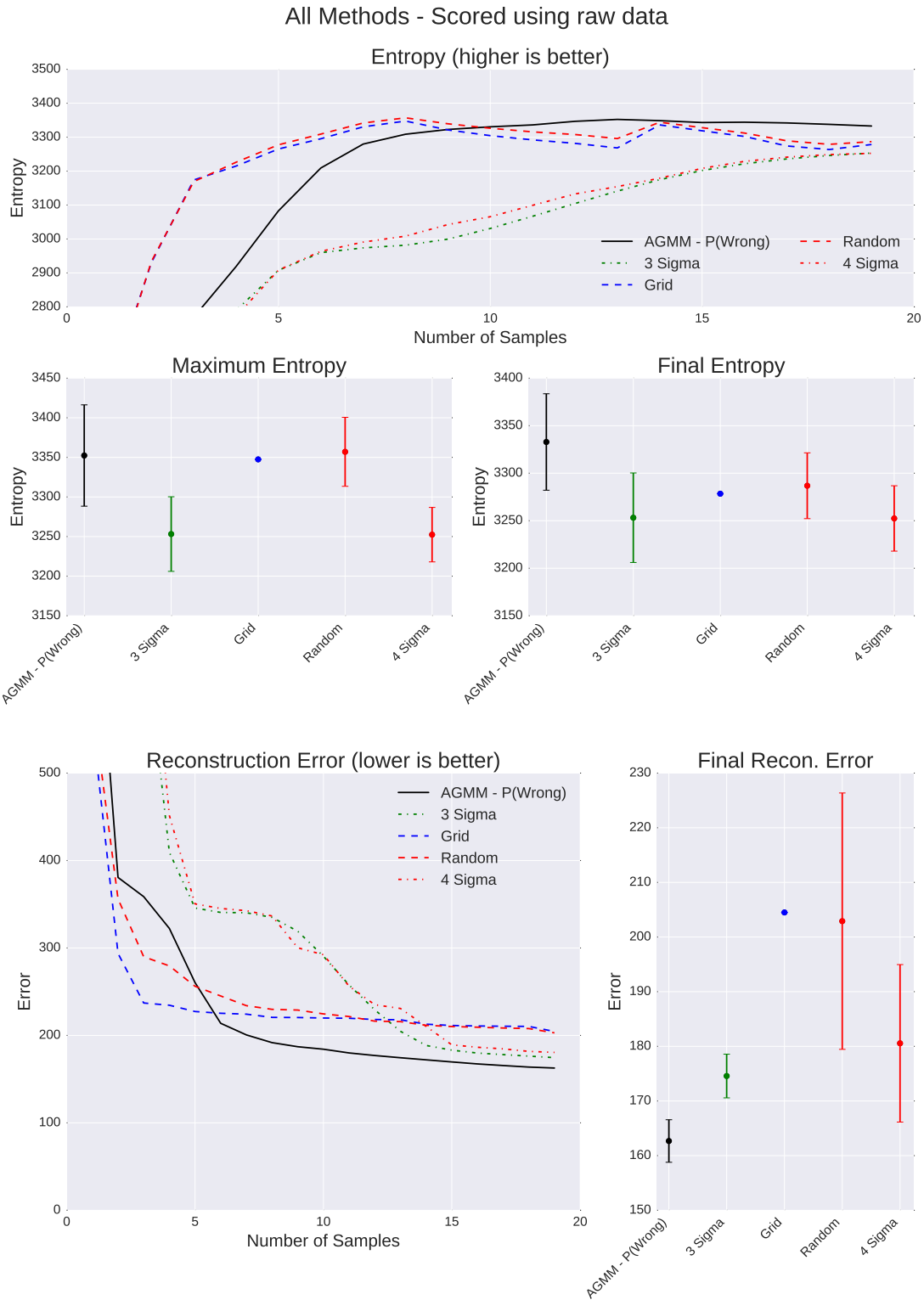


Figure 6.6: Entropy and reconstruction error, scored on the full-channel data.

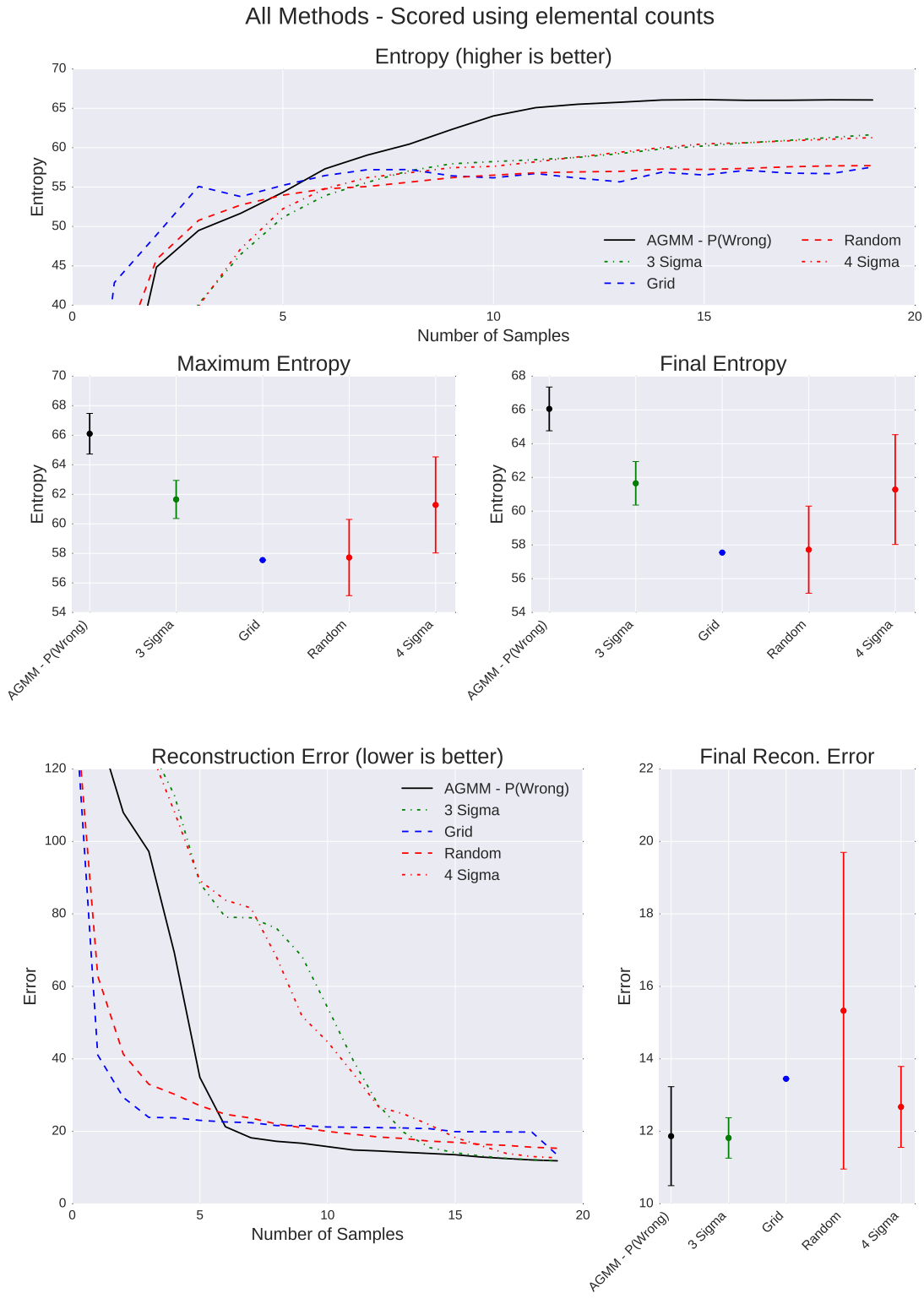


Figure 6.7: Entropy and reconstruction error, scored on the reduced-channel dataset.

6.3.1 Detecting Rare Signatures

We would also like to evaluate these approaches in a manner more similar to how they would be used in a planetary exploration mission, and not just based on the information present in the set of samples.

As an alternative to the entropy and reconstruction error metrics, each method is evaluated based on their ability to detect rare signatures present in the data. This is done in the same manner as in Section 5.3.5, where the full-context approaches were evaluated on their ability to sample rare and uncommon elements.

Given the same abundance maps that were discussed in Section 5.3.5, the number of elements present in each sample, as well as their rarity, is calculated. Rarity is defined as the percentage of the image that has each element present. We then group elements into three rarity classes: Common (present in 10-100% of the image), Uncommon (1-10% of the image), and Rare ($< 1\%$ of the image).

Method Name	Uncommon Elements	Rare Elements
Grid	P, Pr, C, Y, Bi	Pd
Random	K, P, Pr, C, Y, Bi, Be	Zr
3Sigma	P, Pr, C, Y, Bi, Be	Zr
4Sigma	K, P, Pr, C, Y, Bi, Be	Zr
AGMM-P(Wrong)	P, Pr, C, Y, Bi, Be	Zr

Table 6.3: Rare elements detected by each approach. Rare elements are those that appear in less than 1% of the test data.

Under this scoring metric, **3Sigma** scores highest, sampling the greatest number of uncommon and rare elements. **AGMM-P(Wrong)** is the second highest-performing method with regards to both uncommon and rare elements.

These results indicate that the interval-based approaches are very good at detecting rare elements. Because they choose to sample when any one of the elements observed does not fit well to the mean, as rare signatures are observed they are often sampled, as the mean value for these elements will be close to zero.

The **AGMM-P(Wrong)** method likely samples some of these points as well, but future points that share a similar value are likely to be passed over, while they would be sampled by an interval-based approach. These points are passed over because they are similar to the prior sample that was collected, resulting in a low expected reward. Depending on the sampling budget this behavior may be preferred, as the interval-based methods are repeatedly sampling similar points.

Table 6.3 shows the rare elements detected by each method, and Figure 6.8 shows the percent of samples in each method that sample an uncommon or rare element.

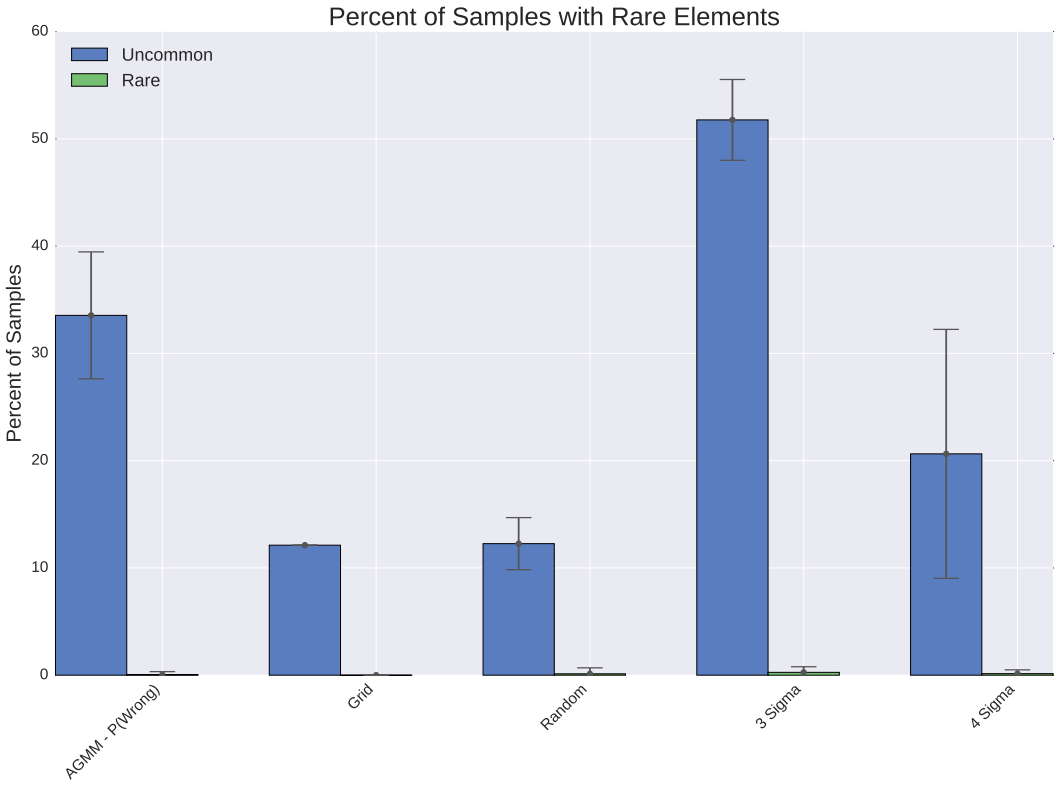


Figure 6.8: The percent of samples that contained Uncommon (blue) and Rare (green) elements, on average, by each approach. Approaches are ordered from left to right based on their entropy scores in the previous section, in increasing order.

Chapter 7

Adaptive Path Planning

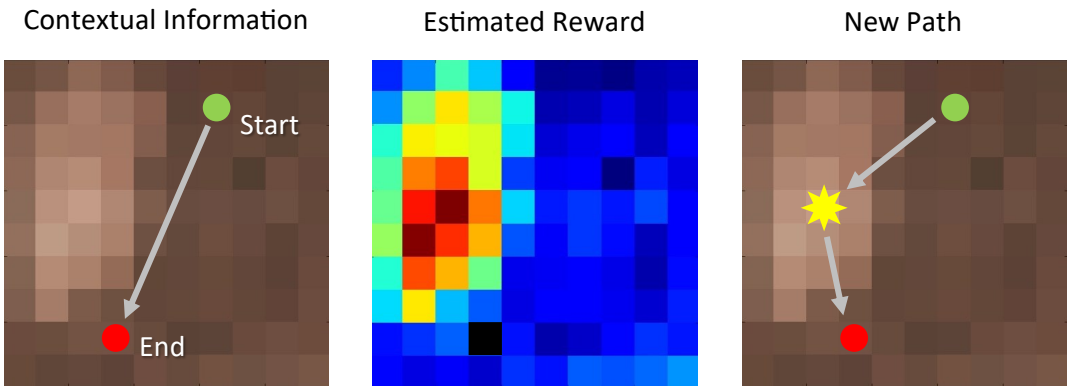
An important part of rover operations is not just pointing instruments and collecting samples, but choosing where the rover physically goes in its traverse. This is commonly done by a team of scientists, such as the Science Operations Working Group, the team who directs the MER and MSL rovers [NASA, 2016c].

Currently, it is possible for a team to manage the day-to-day planning and operations of rovers, and, due to the cost and risk aversion, this will potentially be the case for the near future. However, as rovers become more abundant, more reliable, and more autonomous, and as planetary exploration techniques become more widespread on terrestrial platforms, it is unsustainable to have teams micromanage every rover operation.

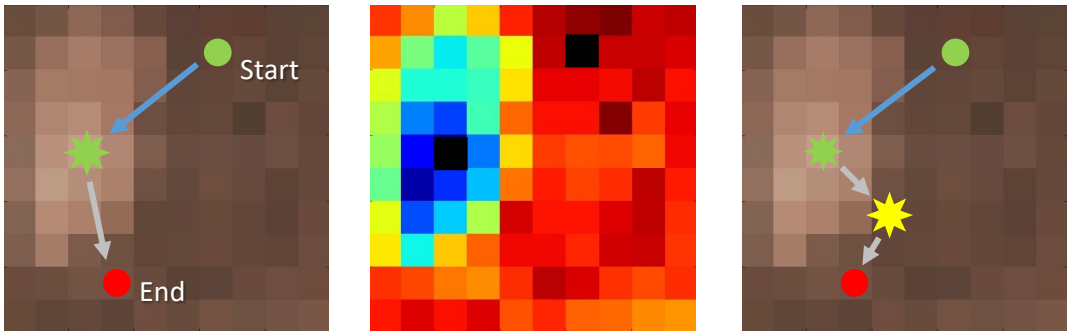
In this series of experiments we will explore techniques for improving a rover's ability to plan a sampling path and adapt to measurements as it progresses along that path. An example of the process is shown in Figure 7.1. This is an interesting extension of the prior chapters, as it offers a number of additional constraints. First, there are constraints on both the length of the path and a cost associated with sampling at each point, as opposed to our prior tests where there were only limitations on the number of samples selected. This limits the planning horizon, and introduces limitations in some of the methods, discussed in more detail in Section 7.3.

Second, methods are now planning over an entire path, not just one step ahead. This is a more difficult problem, as the error present in each expected reward compounds as the length of the path grows.

Finally, as in the other experiments, sampling at a location fully reveals the data at that point. This leads to a change in the reward of the precomputed path, often necessitating large changes in measurement locations. Thus, the best methods are those that can most accurately predict future reward.



(a) Choosing the first sample point



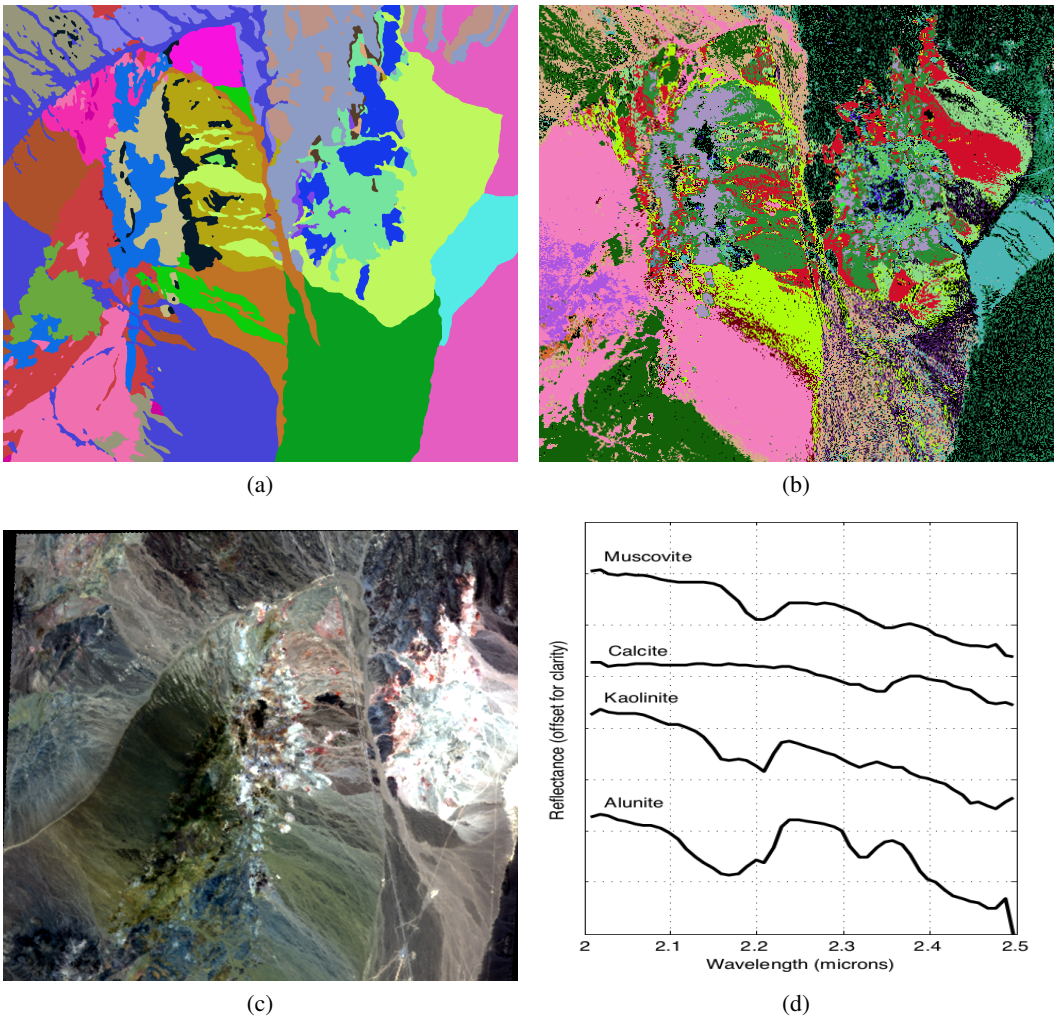
(b) Choosing the second sample point after traveling to the new location

Figure 7.1: An example of the sampling process in the orbital case. Left column: The low resolution context is observed. Center: Rewards are estimated given prior samples. Right: A new path is planned that maximizes the expected joint reward of all sampling points.

7.1 Combining Orbital and Rover Measurements

We design the experiment to simulate a Martian exploration mission, in which a rover collects measurements in a traversal across simulated planetary terrain. The ground data used for this experiment is designed to be similar to that collected by a rover, and the orbital data is similar to that of the Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) instrument [Murchie et al., 2007].

We test this set of experiments on data collected from the Cuprite region of Nevada. Cuprite is a region with a very compact and diverse geologic history. Extensive field tests have been performed at the site, leading to rich geologic and mineralogic maps like those shown in Figure 7.2 [Swayze et al., 2014].



(a) (b) (c) (d)

Figure 7.2: Cuprite, NV. Left: a geologic map of the region. Center Left: mineral abundances. Center Right: a visual image of the region, as taken by the AVIRIS-NG sensor. Right: spectra representative of key minerals in the region. From [Swayze et al., 2014].

Our experiments focus on aligning orbital measurements with those collected by the rover. In the absence of rover-collected measurements of Cuprite, we use imagery from the Next Generation Airborne Visible Infrared Imaging Spectrometer (AVIRIS-NG) [Green et al., 1998, Hamlin et al., 2011]. This instrument has mapped the Cuprite region at high resolution (3.9 meters per pixel), over the 426 wavelengths ranging from 380 to 2510 nanometers. The data were acquired during flights in 2014, and have been converted from radiance to reflectance measurements using the approach outlined in [Thompson et al., 2015c].

We use this AVIRIS-NG data as a proxy for rover measurements, as it is high resolution, both

spatially and spectrally. It is similar to measurements taken during our own experiments in the Atacama Desert, Chile, using the Zoë robot. The spectrometer used on Zoë has a similar range, but with a slightly higher 1 nanometer resolution. While the spatial resolution is not as compact as the measurements collected by Zoë in the Atacama Desert, potentially as dense as centimeters apart, it offers spatial coverage that is difficult to match with a rover.

The orbital data we compare against is generated from the higher-resolution AVIRIS-NG imagery. While the spatial resolution of CRISM is 18.4 meters per pixel at a nominal height of 300 kilometers, we create our mock CRISM data by subsampling our AVIRIS-NG image at a 5:1 ratio, giving our CRISM imagery a spatial resolution of 19.5 meters per pixel.

The spectral range of the CRISM instrument is 362-3920 at 6.55 nanometers per channel, which fully overlaps the spectral range of the AVIRIS-NG data. To avoid complicated mappings between simulated rover and orbital data, we keep the same AVIRIS-NG channels in our simulated CRISM image, just augmented with noise. The noise added to the system is designed to match that specified by the instrument team in [Bultel et al., 2015].

The resulting dataset is a high-resolution and low-resolution pair of images. The high-resolution AVIRIS-NG image is five times the resolution of the CRISM image, so there is a simple mapping of points between the two images.

7.2 Adaptive Methods for Path Planning

We consider a number of adaptive sampling methods for orbital path planning. In the experiment that allowed approaches to look at the full set of contextual information, discussed in Chapter 5, a number of methods were found that perform well in adaptive sampling scenarios, and all of them apply here. In fact, the only requirement for our formulation here is that all methods (aside from the **Random** and **Direct** baseline methods), provide some function over the expected reward of any candidate sample site. Then, as discussed in more detail in Section 7.2.1, points are selected that offer the greatest expected joint reward that are within our sampling budget.

While we exhaustively covered a number of sampling techniques in Chapter 5, we only consider a subset of the best-performing methods here. We will discuss the methods and implementation details in the sections below.

7.2.1 Path Generation

Ideally, we would like to calculate the path with the subset of measurement locations that produce the maximum joint reward. However, this is computationally intractable for large scenes. Instead, we use a simpler method to create a path given the expected rewards at candidate sampling sites.

Initially, we begin with only a start point and an endpoint. At each iteration we compute the expected reward of a potential sampling point, given all data collected prior:

$$r_{n+1} = R(x_{n+1}|x_1, ..., x_n) \tag{7.1}$$

We then sort all points by their rewards, greatest to smallest, and iterate through the list of points until we find a measurement location that can be visited without exceeding our budget. This point is then added to the candidate path, by inserting it at the location that minimally increases the path length.

Adding a measurement location inherently changes the reward at later measurement locations, so the reward function is recalculated at all unsampled points. The scores are again sorted, and the point with the largest reward that can be sampled within our budget is added to the path.

This selection routine is simple but raises one main concern: points are not guaranteed to be sampled in the optimal order, or even the order they were added to the path. This means that the path chosen might suboptimally sample the terrain, in the fact that the first point visited might not have a high expected reward and sampling that point might cause the rover to replan and deviate from the original path that was estimated to be optimal.

While this is not ideal, this method does also discourage poor path planning behavior. As we will discuss in Section 7.3, if we constrained the path to sample in the order the path was calculated, the rover runs the risk of using most or all of its budget to visit one or two points far from the starting location that are expected to have a high reward. In this case, the total number of points sampled is very low, as the rover uses up most or all of its budget driving to and sampling that single point. This is undesirable behavior, as real planetary missions would be more likely to encourage a larger sample set with potentially lower reward, over the risk of sampling a single point with a high expected reward.

We consider the behavior of sampling multiple suboptimal points to be more desirable than the forcing the rover to sample point strictly in order of expected reward. However, as we will see in the results at the end of this chapter, this behavior of first sampling a distant but high reward target is still present in some tests, and leads to poor results.

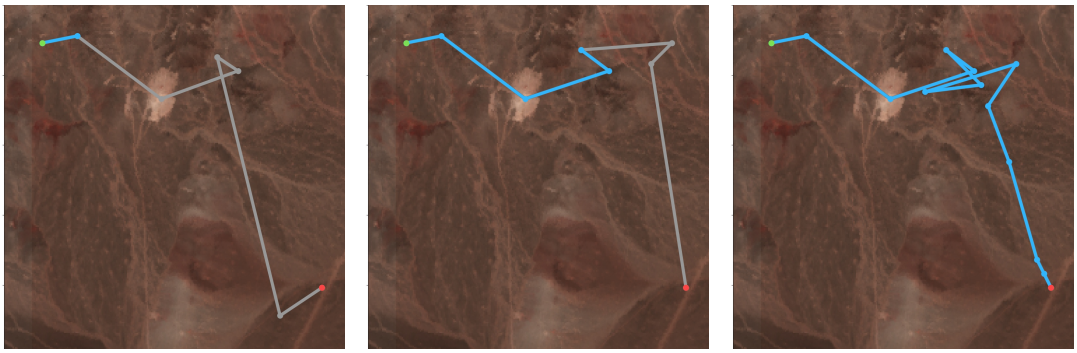
7.2.2 Adaptive Gaussian Mixture Models

We elect to compare all of our AGMM approaches in the orbital path planning: **AGMM**, **AGMM-AddClasses**, and **AGMM-P(Wrong)**.

For both the **AGMM** and **AGMM-P(Wrong)** tests, K, the upper limit of internal mixtures in the model, is set equal to ten. As was demonstrated in Figure 5.9, there is slight benefit to having the number of clusters less than the total number of expected sample points, so this value is chosen as we expect to see an average of 15 to 20 samples collected along each path.

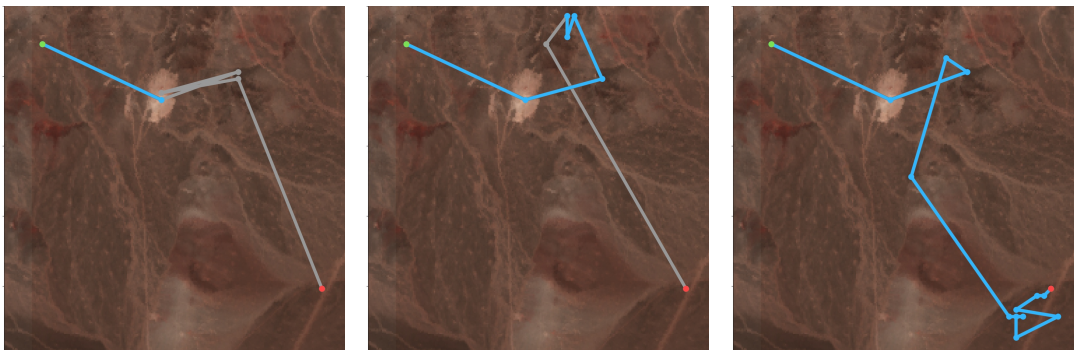
As in the general case, in the **AGMM-P(Wrong)** method we select the point with the greatest probability of misclassification, given our existing model. In the **AGMM** and **AGMM-AddClasses** methods, we choose samples by the finding the sample, within our budget, with the largest Mahalanobis distance from its closest mixture. In the **AGMM-AddClasses** case, we then calculate the probability of misclassification of this point, weighted by the Dirichlet process prior on this data point being from an unknown and unmodeled class, then create a new mixture component if the probability of misclassification is greater than 0.997.

Examples of the different AGMM approaches are shown in Figures 7.3 and 7.4.



(a) Initially planned path (b) Partially sampled path (c) Final path

Figure 7.3: An example of a sample path selected by the **AGMM** algorithm.



(a) Initially planned path (b) Partially sampled path (c) Final path

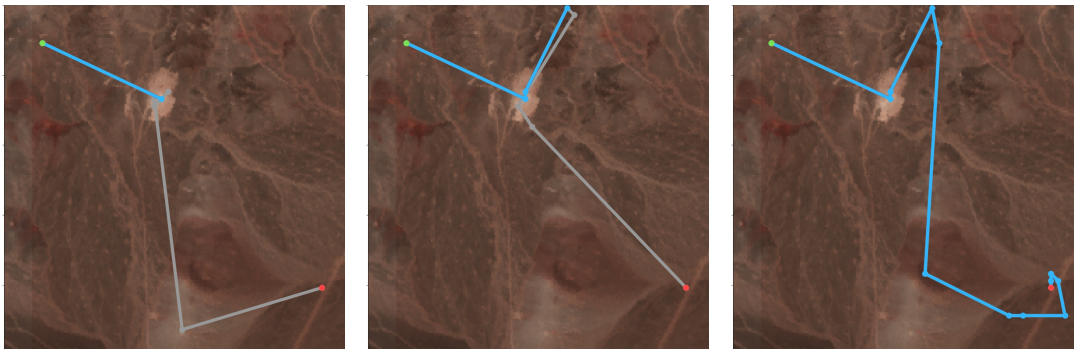
Figure 7.4: An example of a sample path selected by the **AGMM-AddClasses** algorithm.

7.2.3 Bayesian Optimization-Based Methods

We elect to test two Bayesian optimization methods in the orbital path planning tests: **GP-EI** and **GP-MaxMean**. **GP-EI** was shown in Chapter 5 to be a very variable method, depending on the dataset. It has the highest overall entropy score when scored against the raw data, yet is in the middle of the pack when scored against the entropy of the post-processed data, and performs poorly with regards to reconstruction error. Because it did so well on some of the tests, we elect to test it in this scenario as well.

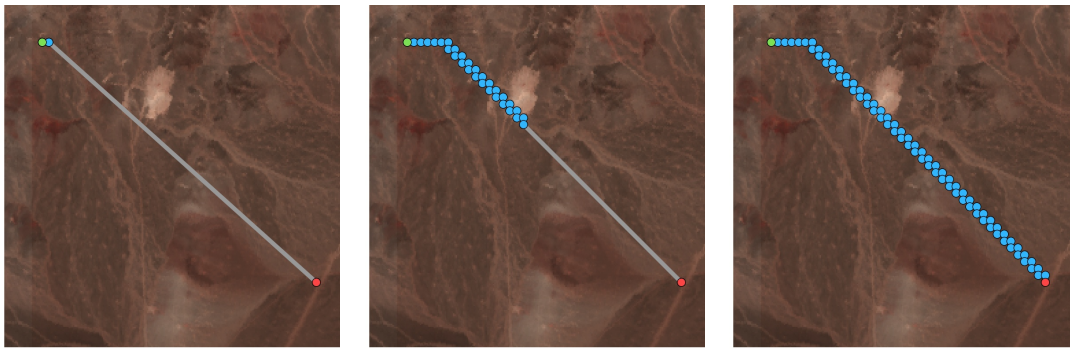
GP - MaxMean is a simple method, choosing the point with the highest expected entropy, as estimated by a Gaussian process. The Gaussian process acts as effectively a smoothing of the data, which was shown to be surprisingly effective in the full context case. **GP-MaxMean** outperformed the baseline methods on all tests, and is the method with the lowest reconstruction error in all tests.

Both of these methods are trained by calculating the expected entropy score at all unsampled points, conditioned on the data already sampled by the rover and the data expected to be sampled



(a) Initially planned path (b) Partially sampled path (c) Final path

Figure 7.5: An example of a sample path selected by the **GP-EI** algorithm.



(a) Initially planned path (b) Partially sampled path (c) Final path

Figure 7.6: An example of a sample path selected by the **Direct** algorithm.

on the path:

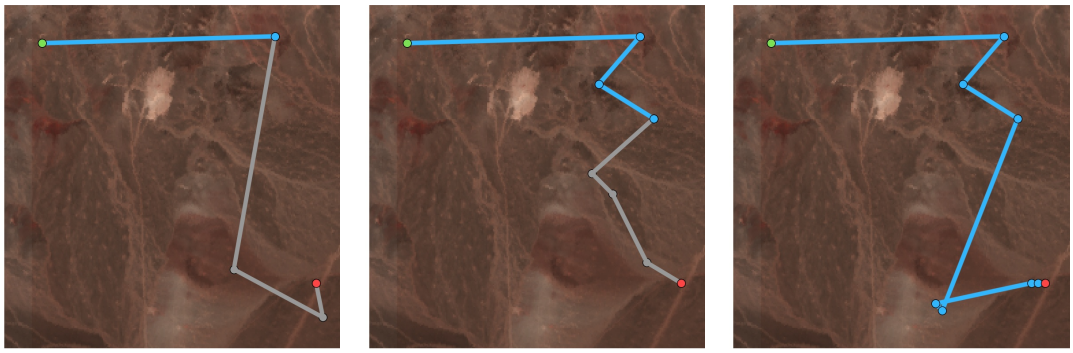
$$r(x_i) = H_{GP}(x_i | X_{sampled} \cup X_{expected}) \tag{7.2}$$

As in the general case, we train the Gaussian process using a random selection of unsampled locations. Similarly to the general case, we use a value of $\xi = 0.01$ for the **GP-EI** method.

7.2.4 Baseline Methods

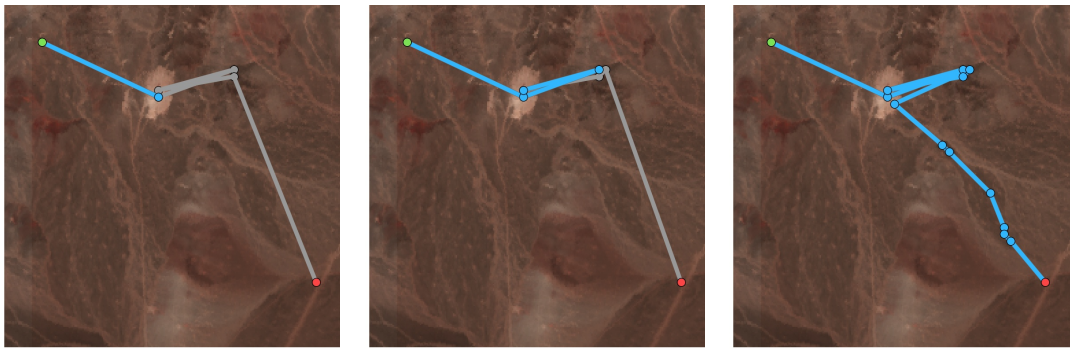
We choose three methods for our baseline comparisons. First, we consider the **Direct** case, in which the rover moves directly from its location to the goal, sampling at points directly along the way. This periodic method of sampling is often used to produce an unbiased and representative dataset of the traverse, as discussed in Chapter 4.

Second, we consider a **Random** sampler. This sampler randomly selects unsampled locations and adds it to its sampling path, until no more locations within its budget can be found. An example



(a) Initially planned path (b) Partially sampled path (c) Final path

Figure 7.7: An example of a sample path selected by the **Random** algorithm.



(a) Initially planned path (b) Partially sampled path (c) Final path

Figure 7.8: An example of a sample path selected by the **MaxEntropy** algorithm.

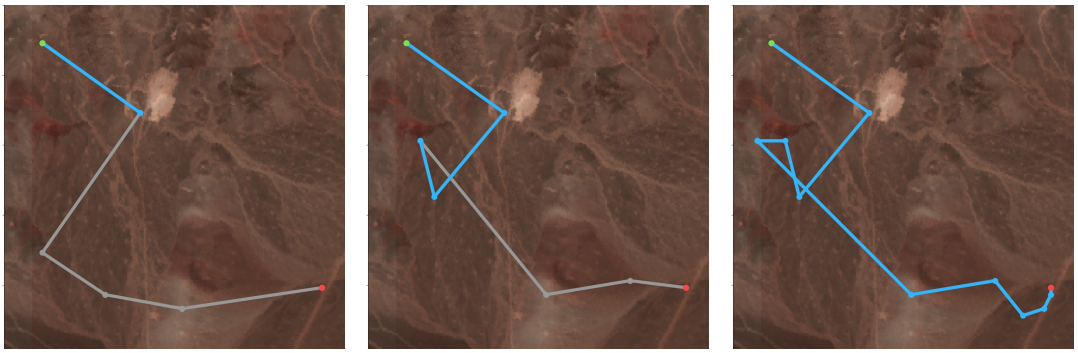
is shown in Figure 7.7.

Third, we consider the greedy entropy maximization algorithm we tested against in prior chapters, **MaxEntropy**. This method scores points based on the differential entropy of their contextual data, given all points currently in the path, both sampled and expected to be sampled:

$$H(x_i|\text{path}) = \frac{1}{2} \sum_{b \in Bands} \log 2\pi\sigma_b \tag{7.3}$$

At every iteration, this approach adds the point that has the greatest expected entropy that can be sampled without exceeding our budget. An example of a path generated using this approach is shown in Figure 7.8.

Finally, we compare against the state-of-the-art method, **SpectroMapper** [Thompson et al., 2015b], discussed in more detail in Section 2.2.1. In this approach, the aim is to sample at the locations that are expected to best reconstruct the orbital spectral image.



(a) Initially planned path (b) Partially sampled path (c) Final path

Figure 7.9: An example of a sample path selected by the **SpectroMapper** algorithm. Unlike the entropy-based selection strategies, **SpectroMapper** avoids the lighter playa in the top center of the image, and instead focuses on more representative measurement locations.

It does this by evaluating a set of candidate sampling sites in the orbital imagery, $\mathcal{L}$, and selecting the subset of locations $B = \{b : b \in \mathcal{L}\}$ such that the spectral library generated by samples collected at B, Y_B, minimizes the expected reconstruction error of the entire remote image:

$$R(B) = E\left[\sum_{\mathbf{x}\in X} \min_{\phi} ||Y_B\phi - \mathbf{x}||_2\right] \tag{7.4}$$

for a non-negative set of mixing fractions ϕ and subject to some overall budget $C(B) < \beta$. As we have not collected samples on the ground, Y_B is approximated by orbital data, X_B.

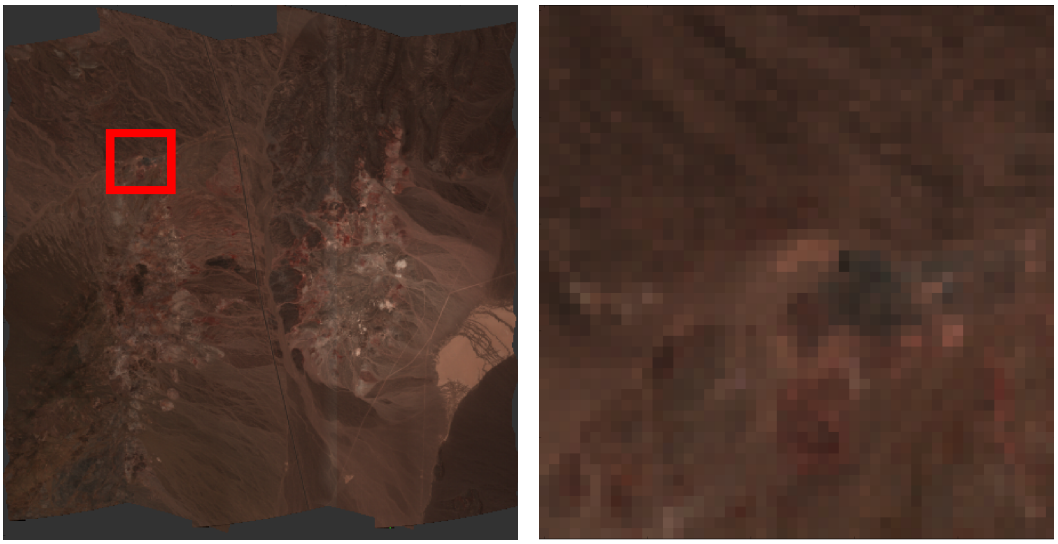
As samples are collected along the path, the collected spectra, Z_A, are used as more accurate data for endmember calculation. The equation then becomes

$$R(B|A) = \sum_{\mathbf{x}\in X} \min_{\phi} ||[Z_A\ X_B]\phi - \mathbf{x}||_2 \tag{7.5}$$

again for $\phi > 0$ and respective of a total budget $C(A) + C(B) < \beta$. An example of the results of this approach is shown in Figure 7.9.

7.2.5 Testing Methodology

Approach are tested on the data described in Section 7.1. The original data is approximately 100 square kilometers, with nearly six million measurements. As this is computationally infeasible for many of our methods, we instead break up the original image into ten 1 km^2 test images, randomly selected throughout the Cuprite region. This distance was selected both to keep the computational requirements reasonable and as this was of similar length to traverses taken during our prior experiments in the Atacama Desert in Chile.



(a) (b)

Figure 7.10: A randomly-selected test region. Left: original AVIRIS-NG image with the selected region highlighted. Right: a selected test region approximately 1km square, at CRISM resolution.

We then select five pairs of start and end locations, which are commonly used throughout all 10 images. Each start and end point pair covers a slightly different region of the scene, and are designed to force each method to travel through all regions of the scenes.

This results in a total of 50 paths tested for each method. In each test, each method is given a budget of twice the length from the start to the end locations. While this is higher than what we used in trials in the field, typically set to 1.2 or 1.4 times the distance of the path due to time limitations, we do not think this is an unreasonable value and it results in paths of longer length, which are interesting for analysis.

During the planning process, the rover chooses measurement locations within the low resolution CRISM imagery. As the AVIRIS imagery, our proxy for rover-scale measurements, is higher resolution, we must decide on a sampling routine to choose which pixels are sampled from the higher resolution imagery within each stop in the low resolution image. More specifically, the AVIRIS image is five times the resolution of the CRISM image, so we must choose a subset of the 25 AVIRIS points within the target CRISM pixel.

We choose to do this by determining the vector from the current rover position to the center of the CRISM pixel. We then determine which points within the 25 AVIRIS pixels we will pass, and sample from each of those, up to and including the pixel at the center of the targeted CRISM pixel. We believe this accurately represents the idea that rover-based measurements are fairly inexpensive, and, while we cannot afford to sample every point along the vector from the start to the end location, it is not prohibitive to sample the points nearest the goal. An example is shown in Figure 7.11.

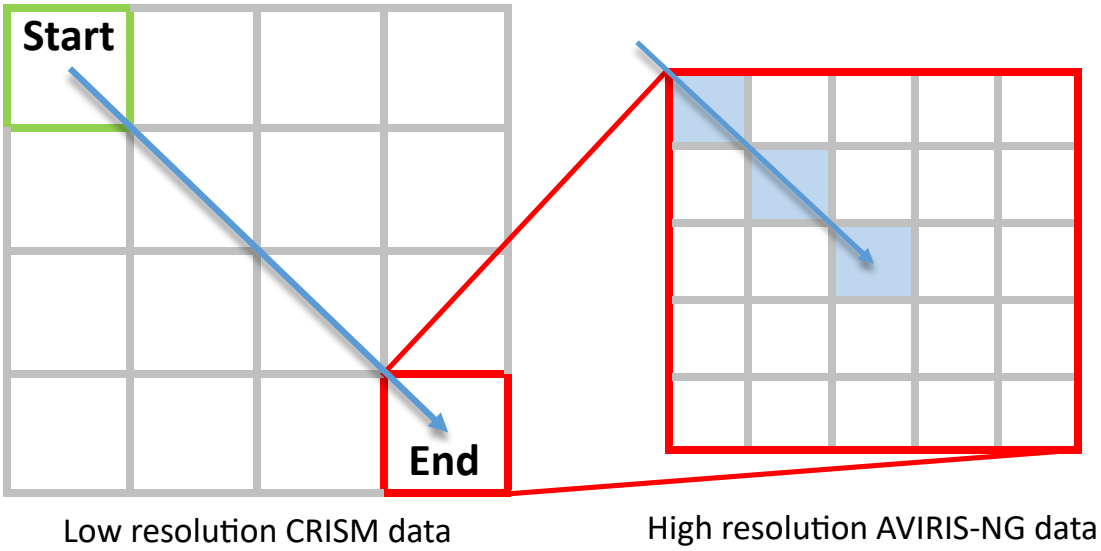


Figure 7.11: An example of how sample points are chosen within the high resolution data. On the left we show the path that is planned in the low resolution CRISM data. When the rover enters the cell containing the measurement location, all points in the high resolution data that are intersected by the path are sampled. This simulates a mission in which samples are cheap relative to the cost of driving, such as a rover collecting quick spectra along its drive.

Finally, in each trial we include cost for sampling, set at 100 meters of the budget. This acts as a deterrent to methods that want to include an excessive number of samples, decreasing their available budget slightly each time they sample a point.

In our original experiments, this cost was not present. As a result a number of methods would select horizontal or vertical paths, then plan stops at every location on that line, as there was no cost per sampling stop, only some potential increase in the expected reward of the overall path. This value is selected as it is large enough to deter paths with low-cost and low-reward sampling stops, yet not high enough to reduce the number of stops to a prohibitively small number.

7.3 Results and Discussion

We find that the **MaxEntropy** method performs very well, outscoring all other approaches with regards to the entropy metric. This is likely because the noise model used to generate the CRISM dataset is, in effect, not noisy enough. Without the presence of enough noise, the Max Entropy method is able to effectively select the best sample set, with regards to the entropy scoring metric.

In real world operations, this is less likely to be the case. Measurements from the rover are

unlikely to exactly match the spectral resolution of the orbital instrument, inducing some noise in the conversion between the two. Furthermore, rovers typically sample in a point-wise manner, meaning their instruments have a small field of view. Samples collected are then more likely to be of a single material, like a rock or ground cover, while the orbital measurements are an integration over the entire region. This causes further discrepancies between the rover and orbital models that would lead to poorer results from an approach like **MaxEntropy**, which relies on accurate contextual data to sample well.

Aside from **MaxEntropy**, the **AGMM**, **AGMM-AddClasses**, and **GP-MaxMean** methods perform very well with regards to the entropy metric. Each of these methods achieves a strong initial score in the first few measurement locations, similar to that of **MaxEntropy**.

With regards to reconstruction error, the **SpectroMapper** and **Direct** methods achieve the best results. The **Direct** method scores well because it samples across the data, producing unbiased and representative sampling. It is no surprise that **SpectroMapper** scores well with regards to reconstruction error, as that is the objective of the approach.

The method out of our proposed approaches that scores the lowest reconstruction error is that of **GP-MaxMean**. These results, combined with the results of our baseline tests in Section 5.3.4, lead us to believe that **GP-MaxMean** is perhaps the best overall method out of our proposed approaches. It scores well in nearly all scenarios, with regards to both entropy and reconstruction error.

An interesting note is that entropy is a somewhat problematic way to score these paths. As can be seen in the gradual curve present in our entropy graphs, the entropy score typically peaks, then gradually declines as redundant information is collected. As a result, methods that select more point, eg. the **Direct** method, are at a disadvantage when comparing the final entropy scores. For accurate comparisons, we consider only the first 25 points when calculating the final scores. Most methods are at or below this value, so it has little impact on anything except the baseline **Direct** method.

In real experiments, however, if the budget allows, it is always better to collect more samples. Even though it may be detrimental to the overall final entropy score, sampling more often presents more opportunities for detection of unknown signatures, or even a more representative sample of something collected at a prior location. If there is a concern with oversampling, then a better strategy might be to sample as often as the budget allows, then compute the most diverse set of n samples, where n is some target number of signatures to return for analysis.

Figure 7.14 shows the results for the orbital tests. All plots are extended to a length of 40 samples to make the results comparable, repeating the final value if they did not sample that many points.

As a side note, entropy here is a negative value, where it was positive in prior tests. This is largely because of the logarithm operator in the calculation of entropy, and the scale of the data considered. In prior tests, the data consisted of line intensities or pseudointensities, both of which are positive and are very often greater than 1.0. This leads to entropy values that are positive, as the logarithm of a value above 1.0 is positive. In the case of the orbital data used in this experiment, each point is a reflectance in the range 0.0 and 1.0. This leads to negative entropy scores.

Additionally, Figure 7.12 plots the average number of points sampled within the high-resolution

Method Name	Max Entropy	Final Entropy	Recon Error
Direct	-1042.2 ± 157.1	-1039.3 ± 155.8	0.1163 ± 0.0351
Random	-1043.7 ± 135.3	-1043.5 ± 136.2	0.1469 ± 0.0431
Maximum Entropy (MaxEntropy)	-544.5 ± 170.9	-614.8 ± 159.3	0.1442 ± 0.0438
SpectroMapper	-908.9 ± 169.2	-914.9 ± 141.9	**0.1184 ± 0.0288**
Expectation of Improvement (GP - EI (Matern kernel))	-1081.5 ± 164.1	-1080.2 ± 163.1	0.1758 ± 0.0705
Maximum Mean (GP - MaxMean (Poly kernel)	-672.8 ± 234.2	-766.5 ± 182.1	0.1428 ± 0.0400
Adaptive GMM (AGMM)	-597.5 ± 223.5	-675.2 ± 114.1	0.1549 ± 0.0357
AGMM - Add Classes (AGMM - AddClass)	**-592.1 ± 194.7**	**-681.4 ± 110.5**	0.1719 ± 0.0438
AGMM - P(Wrong) (AGMM - P(Wrong))	-955.3 ± 140.3	-961.6 ± 142.0	0.1586 ± 0.0536

Table 7.1: The overall results of our tests on the orbital path planning scenario. Our contributions and the best results, excluding Max Entropy, are listed in bold.

imagery, per path, for each method. We see that, unsurprisingly, the **Direct** approach has a much larger number of samples, as it spends nearly all of its budget on sampling operations instead of travel to measurement locations. Aside from the **Direct** path, we see a high amount of variance in the **GP-EI** method, which is perhaps part of the reason it scores so poorly on these tests.

When investigating the poor performance of **GP-EI**, we notice an interesting phenomenon with many of its tests, in which the first point or first few points selected are far from both the start and then end point, exhausting much of the exploration budget. An example is shown in Figure 7.13. In these cases, the method may initially find a high-scoring point and have planned a path that maximizes the future expected reward with regards to that point, then, upon reaching and sampling the point, finds that the planned path no longer offers a high reward. The rover is then limited in which points it can choose for follow-up, as much of the budget is exhausted, and is forced to travel within a limited corridor from that sample point to the goal, limiting its ability to choose diverse sites. This may also be the reason why there is such high variance in the number of points sampled by **GP-EI**, as it is more prone to get "stuck" in these situations, which can lead to both a very high or very low number of samples per path.

In these cases where the rover exhausts much of its budget sampling the first point, the overall result is typically a poor path score. Potentially we could place limitations on the ability of methods to plan paths like these, or try to further break the path up with measurement locations along the way so these distance points are not the first locations sampled by the rover. We do not explore any of these options in this thesis, but discuss the problem more in Section 8.2.

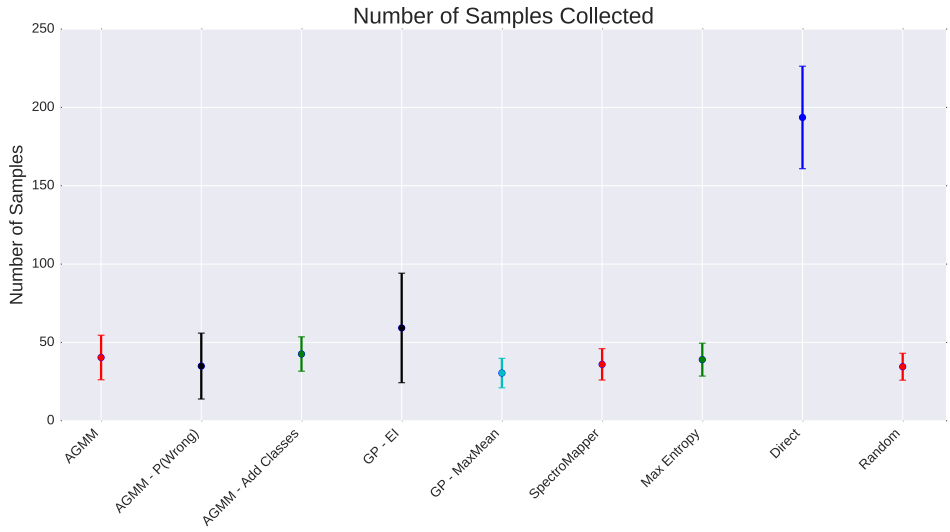
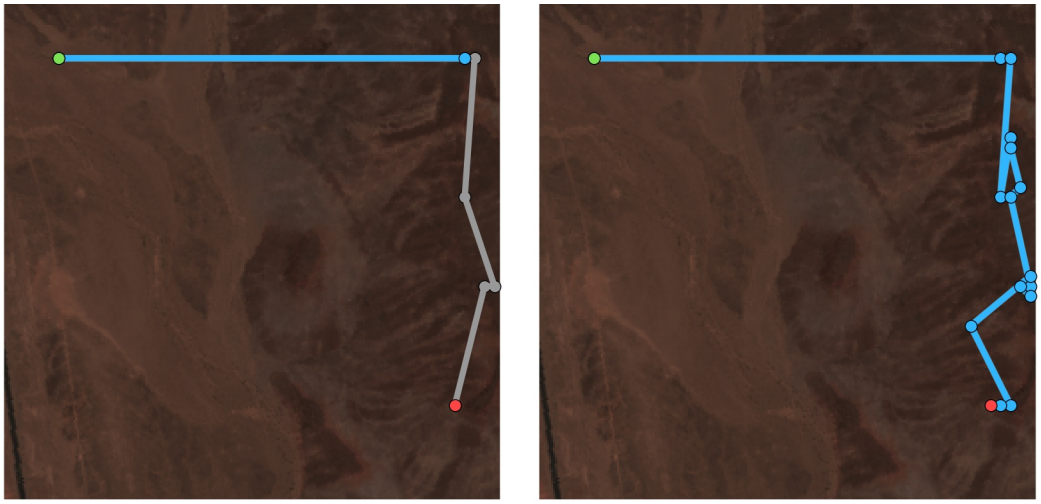


Figure 7.12: Average number of samples collected per path for each method tested.



(a) Initially planned path

(b) Final path

Figure 7.13: An example of a poorly-planned path that uses much of its initial budget sampling the first location. The rest of the path is then constrained to the far right side of the map. A more reasonable path might sample along the way to the first location, improving its understanding about the terrain without expending so much of its budget. The rover is constrained to start at the green point and end at the red point.

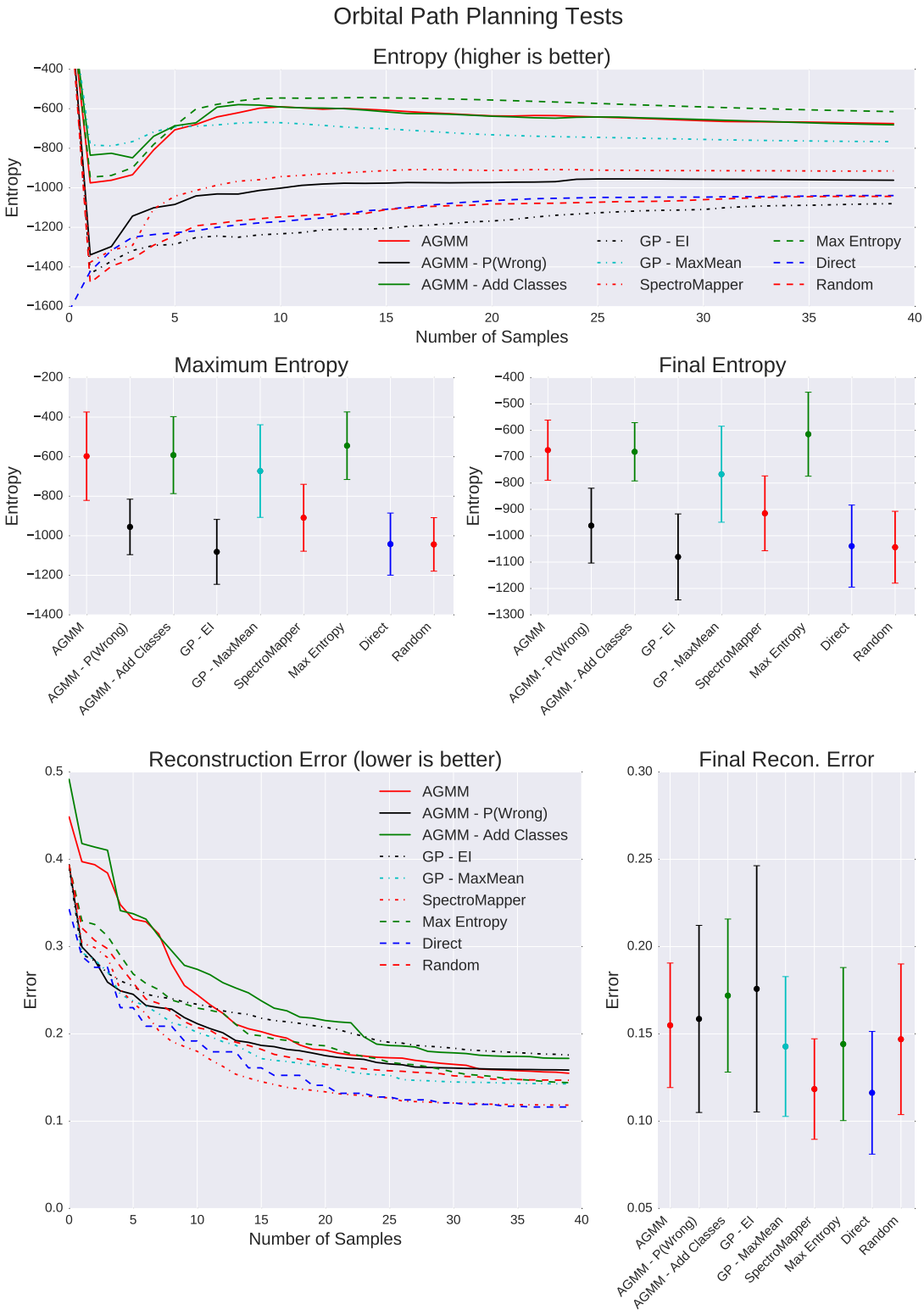


Figure 7.14: Entropy and reconstruction error, scored on the AVIRIS-NG data.

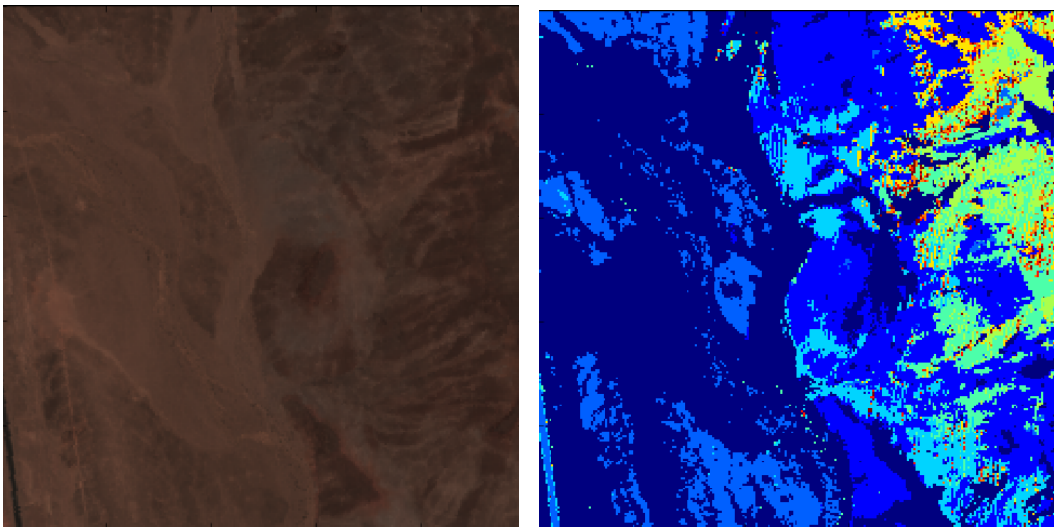


Figure 7.15: Left: One of the test regions for the path planning experiment. Right: The PCA segmentation of this region, where different colors indicate different class labels.

7.3.1 Sampling From Representative Terrain Classes

As in the prior experiments, we also evaluate these approaches in a manner more typical of exploration robotics missions. Planetary missions, such as the Mars 2020 mission, are often concerned with collecting observations at a variety of different locales or terrain classes, as these regions are likely to contain different geology or terrain features that might be of interest to planetary scientists and astrobiologists.

To emulate this, we evaluate the number of terrain classes that are visited, on average, by each approach. As the orbital map does not contain class information, it is segmented using Principal Component Analysis (PCA) [Jolliffe, 2002].

PCA produces a number of principal components, as well as weightings for each component at each location. The principal components are constructed such that they are an orthogonal basis set for the data, and the weightings indicate the contribution of each principal component to the original data.

To cluster using this data we compute the principal components and the weightings, then select the class as the largest-weighted component at each location. This produces a segmentation like that shown in Figure 7.15.

Each approach is evaluated on the number of distinct classes visited. Because some approaches, such as the **Direct** approach, sample many more times than other approaches, this score is normalized by the number of sample points collected.

From the results, shown in Figure 7.16, **GP-MaxMean**, **SpectroMapper**, and **Random** all do very well. Each of these approaches visits a high number of new terrain classes, relative to other approaches.

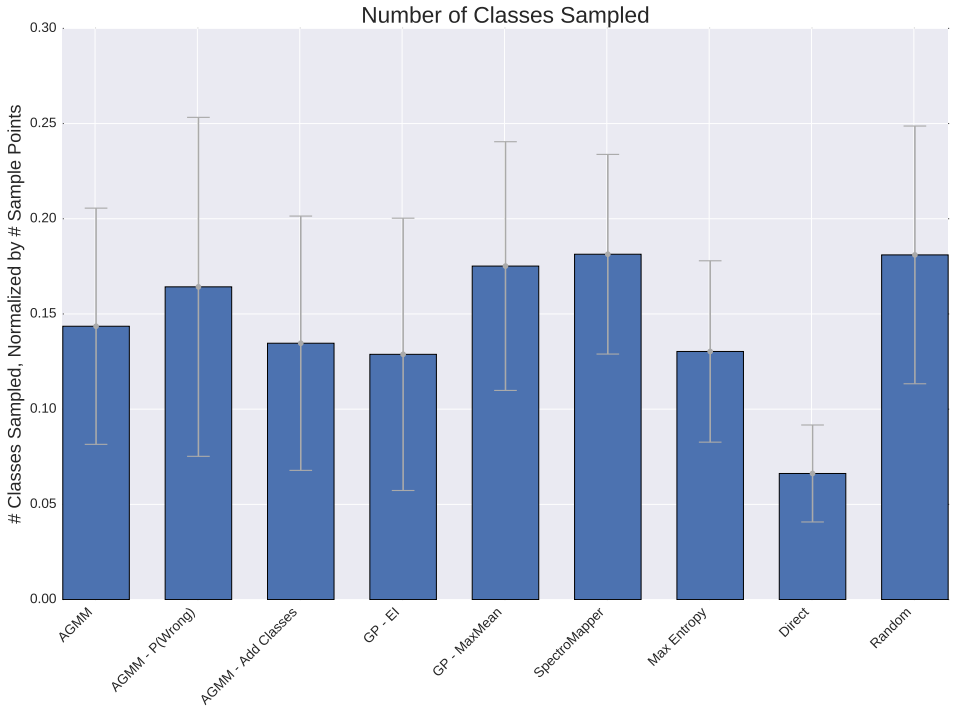


Figure 7.16: The number of classes sampled for each approach, normalized by the number of sample points in each approach's path.

Out of the AGMM-based approaches, **AGMM-P(Wrong)** scores the highest. Given that the P(Wrong) metric explicitly calculates the misclassification probability, this is unsurprising.

Chapter 8

Conclusion

This research has advocated that adaptive sampling methods increase a robot's ability to observe its environment, collect samples, and react to that information. The methods shown here improve upon established adaptive sampling techniques, increasing the diversity and representativeness of samples collect. Furthermore, the approaches introduced here are able to better detect and sample the uncommon and rare atomic signatures and terrain classes observed during sampling. These methods perform well in the presence of noise and work at all scales of data, from microscopic to orbital.

8.1 Contributions

We have made significant contributions to the field of adaptive sampling. We have demonstrated improvements to adaptive sampling that apply across all scales of robotic exploration, bringing together concepts from active learning, information theory, and Bayesian optimization. Our contributions are as follows:

- Adaptive Gaussian Mixture Models
  - This approach is more resilient than traditional models when constructed from a low number of samples.
  - We improved upon this by adding a Dirichlet process prior over the size of the model, giving the it flexibility to grow over time and react to discrepancies between the expected value of a sample and the true value of the sample once observed.
  - We also show that Dirichlet process priors can be extended to computing the probability misclassification, providing better misclassification estimates by accounting for unknown and unobserved classes of data.
  - These approaches show consistently large improvements over established adaptive sampling methods in all experiments.

- One-Step Bayesian Optimization
  - Multiple acquisition functions were tested, each offering a trade-off between uncertain points that have the potential for a very high reward or well-understood points expected to have a high reward.
  - These approaches were shown to excel at finding diverse samples and excelled at finding uncommon and rare atomic signatures.
  - **GP-MaxMean** performs especially well, consistently being in one of the top-performing methods, under all metrics.

We evaluated these approaches in three experiments across all scales of robotic exploration, each directly applicable to common sampling scenarios in exploration robotics.

- Adaptive sampling with full context considers a robot selecting informative sampling locations, given contextual data for an entire scene.
- Adaptive sampling with limited context considers a microscopic sampling scenario in which contextual data is observed one at a time and prior sampling locations cannot be revisited.
- Adaptive path planning considers a rover planning a traverse, using orbital data to estimate rewarding sample locations along the path.

Finally, we have considered not just evaluations of the amount of data present in each sample but science-relevant metrics for evaluating different adaptive sampling techniques. These metrics show that our approaches are competitive or improve upon the number of rare or uncommon atomic signatures detected, as well as sample from diverse regions when selecting paths from orbital data.

8.2 Future Work

There are a number of potential improvements to this work and future avenues of investigation. First, an improvement we would like to make to our Bayesian optimization approach is to explicitly model the known Poisson noise in the Gaussian process model. For channels with a large enough Poisson mean, $\lambda_b > 10$, we can assume the noise is Gaussian and additive. Then it is simply a measure of including it in the covariance kernel:

$$K = K + I\sigma_B^2 \tag{8.1}$$

where σ_B^2 is a vector with the noise in each channel. This yields the predictive distribution [Brochu et al., 2010b]:

$$P(x_{n+1}|D_{1:n}, x_{n+1}) = \mathcal{N}(\mu_n(x_{n+1}), \sigma_n^2(x_{n+1}) + \sigma_{noise}^2) \tag{8.2}$$

and the statistics

$$\mu_n(x_{n+1}) = k^T[K + \sigma^2_{noise}I]^{-1}y_{1:n} \quad (8.3)$$

$$\sigma^2_n(x_{n+1}) = k(x_{n+1}, x_{n+1}) - k^T[K + \sigma^2_{noise}I]^{-1}k \quad (8.4)$$

This in turn requires adjustments to the EI and PI acquisition functions discussed in Section 3.2. Instead of using the best observation, we use the distribution of the sample points and define the incumbent, $f(x^+)$, our best point thus far, as the point with the highest expected value:

$$\mu^+ = \arg\max_{x_i \in X} \mu(x_i) \quad (8.5)$$

Furthermore, there has been work on resampling potential points to reduce noise, often called *intensification* [Bartz-Beielstein et al., 2005, Huang et al., 2006, Hutter et al., 2009]. This process is especially applicable in the case of spectral co-adding, the process of taking repeated spectrum of a single target and averaging their results to increase the signal-to-noise ratio. Intensification approaches could be used to model this data as it is acquired, deciding when the co-adding process should stop or if it should terminate early.

As a second improvement, in our work with adaptive path planning using orbital image priors, we would like to explore a more principled use of the uncertainty in each path. More specifically, within our Bayesian optimization methods, we are able to calculate an uncertainty at each point in the path as it is selected. This is used at each step in many of the Bayesian optimization methods, but the uncertainty of the entire path is never considered.

In contrast, we would like to explore modeling the uncertainty along an entire path, as it is planned, then use that uncertainty to choose the path, similar to how many of the Bayesian optimization methods use uncertainty in their acquisition functions. This uncertainty could be used similarly to the **GP-UCB** method, in which the upper bound of a path is calculated, in which case the path has the potential to have a very high reward, perhaps with the risk of a suboptimal result. Alternatively, paths could be chosen that still score well, but have a lower uncertainty value, giving some guarantee, with high probability, in an expected reward.

Finally, recent advances in portable multispectral imagers present an interesting avenue for this work at the rover scale. Arizona State University has developed the NIR Cam, a hyperspectral camera that captures multiple-megapixel images at nine different wavelengths, from blue to near infrared [Krishnan et al., 2013]. This could provide a rover with full-scene hyperspectral information, vastly improving the potential for adaptive instrument targeting.

8.3 Improving Agricultural Robotics through Adaptive Sampling

Adaptive sampling is not limited to planetary or exploration robotics, but can also be used to improve a number of terrestrial robotics applications as well. Agricultural robotics is a field that has seen rapid robotics development in recent years, and current investigations have illustrated the need for advanced robotics sampling techniques.

Figure 8.1: Vision algorithms as applied to agricultural robotics. Here crop yield is estimated using a grape detection algorithm [Nuske et al., 2011]. Our approach could be applied to analyze the health of detected grapes to see which vines warranted follow-up.

As the global population increases, a larger and larger productivity demands are placed on on agriculture. By the year 2050, it is predicted that crop production must double in order to keep up with the global population [Tilman, 2011]. In order to meet this demand, crop yields must increase at a rate of approximately 2.4% per year, yet the current average rate is only 1.3% [Ray et al., 2012].

However, over the past 50 years there have been extensive efforts to increase productivity and develop methods for analyzing and improving crop yields. Cereal production has tripled with only a 30% increase in land area cultivated, largely through crop research, infrastructure, and market development [Pingali, 2012].

Recent advances in high throughput genotyping and phenotyping are promising methods for further increasing crop yield. These involve rapid analysis of genomic and physical traits of plant specimens, which can then be used to characterize and determine the causes of plant performance. However, collecting this data is onerous, and the data itself is so vast and complicated that advanced machine learning and data processing techniques must be developed to make the best use of it [Cabrera-Bosquet et al., 2012].

Robotics is a key step in the advancement in high throughput phenotyping, specifically in the rate and scale at which measurements are collected. Aerial vehicles, such as quadcopters, can provide maps of crops. Ground vehicles can collect more detailed observations or even physical samples for onboard or laboratory analysis.

Agricultural robotics has focused on creating two main types of platforms: sensor networks that characterize the health of the crop, and robotic platforms used for targeted analysis of specific crop samples. Sensor networks, like those described in [Kohanbash et al., 2011], provide a large amount of data about the crop but provide no active method for sampling. Conversely, robotic platforms provide a method for very detailed sampling using sensors like spectrometers and thermal cameras, yet only within a very localized region. Both approaches are necessary to characterize the health of a crop.

However, high throughput phenotyping is a numbers game; the more environments and breeds

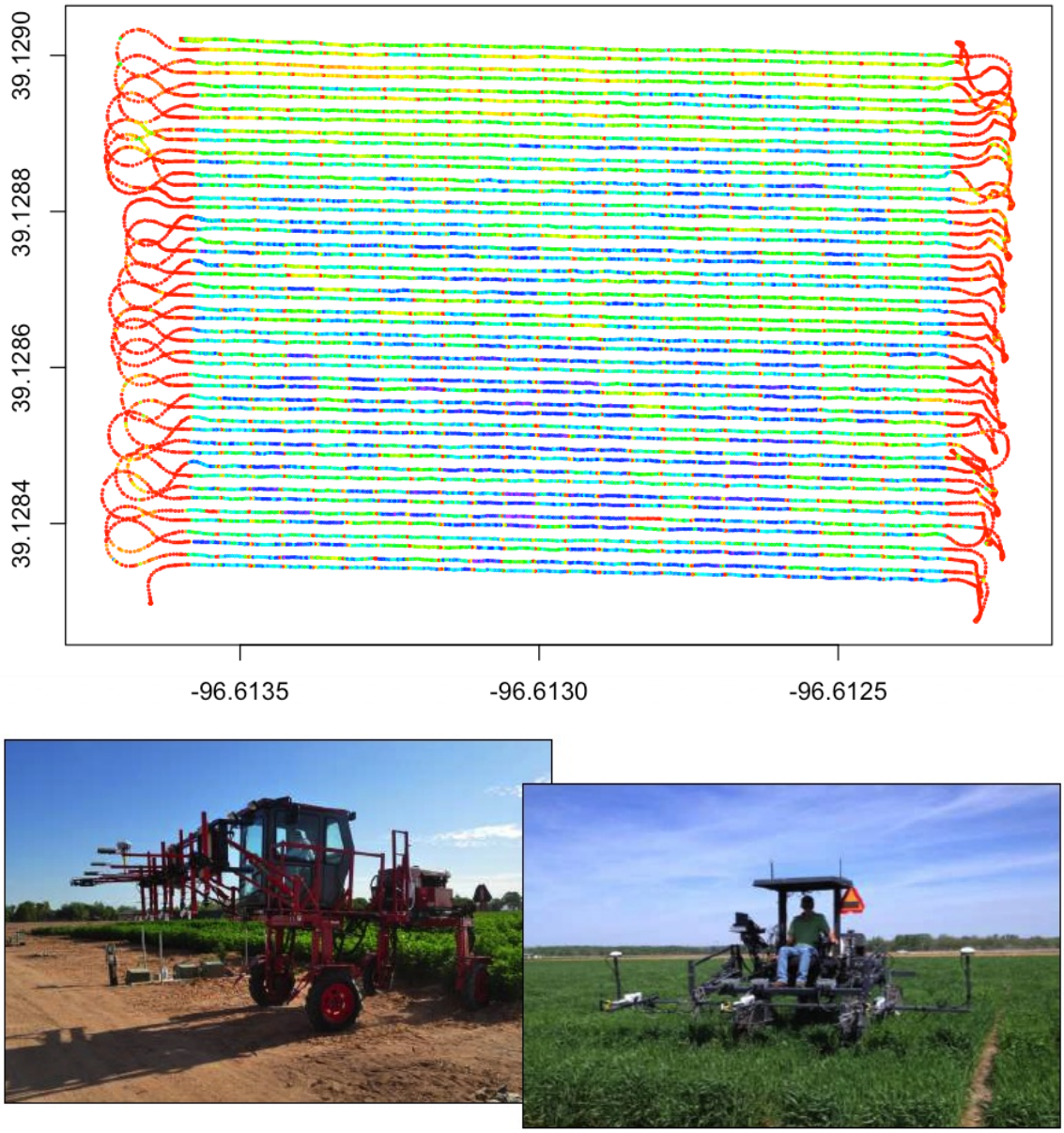


Figure 8.2: Plant phenotyping is a burgeoning field aided by both robotics and machine learning. Top: a quick pass of sensor readings of a field [Poland, 2014]. Readings like these are noisy but can be gathered efficiently and repeated measurements offer valuable temporal context. Bottom: two examples of the machines used to collect these measurements [Poland, 2014]. Our proposed method would enable more accurate follow-up at questionable regions while minimizing time and resource intensive manual collection.

tested, the higher the likelihood of finding superior variations [Araus and Cairns, 2014]. Intense sampling is prohibitively time- and resource-intensive for practical agricultural operations. Instead of approaches like full, regular, random, or human-guided sampling, improvements in adaptive sampling are necessary for characterization of crops using minimal targeted measurements. In the case of sensor networks, combinations of sensed variables, such as pH or soil moisture, can be used as priors to find regions of crops with similar growth patterns or outlier regions that require investigation.

Adaptive sampling methods in agricultural robotics have a number of similarities to the sampling requirements of planetary exploration missions. Specifically, both scenarios deal with multi-scale data, necessitating methods that can direct a robot to promising sampling regions, while also requiring methods that analyze data collected by the robot and determine specific plants or samples most promising for detailed follow-up. Agricultural robots are similarly limited by the amount of time they can explore, as well as the number of samples they can collect, so adaptive methods are necessary to improve their analysis.

Our approach, shown to perform well in planetary scenarios, could also be applied to agricultural robotics in a similar fashion. Aerial or orbital data can be used to select promising measurement locations, using methods like those we proposed in our adaptive path planning experiment. Then, upon observing a site and collecting contextual information, adaptive methods like the ones demonstrated in the full context and limited context experiment could be used to choose locations for the collection of physical samples, or targets for instrument follow-up. Our methods could even be targeted on objects of interest using pre-processing like that proposed by Nuske et al. [Nuske et al., 2011]. Using their grape detection approach, we could provide measurement locations for spectrometer follow-up that would more accurately characterize the health of the cluster.

8.4 Closing Remarks

The capabilities and rate of information acquisition of rovers is rapidly increasing, and it is important that rover intelligence increases at similar rate. Adaptive sampling is one of many ways that rovers can improve the rate at which they sample, the diversity and coverage of the samples they collect, and the overall science return of missions.

We have focused our experiments in this work on the planetary exploration scenario, but these techniques are applicable to terrestrial robotics as well. The number of planetary robotics is growing at a slow enough rate that, for the foreseeable future, teams of scientists micromanaging rover operations is currently sustainable. Except in the case of certain instruments, such as PIXL, this is unlikely to change any time soon.

However, terrestrial robots that are commanded and overseen by teams of humans are no longer the norm, so it is perhaps these scenarios where adaptive sampling is most relevant today. Terrestrial robots are comparatively cheap and easily serviceable, and immediately benefit from rapid improvements in sampling and sensing technology. Thus, these improvements to adaptive sampling improve not just planetary robotics, but robotics as a whole.

Bibliography

[Abramowitz and Stegun, 1965] Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, pages 716–723.

[Aldous, 1985] Aldous, D. (1985). Exchangeability and related topics. *Ecole d'ete de Probabilites de Saint-Flour*, page 1198.

[Araus and Cairns, 2014] Araus, J. L. and Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science*, 19(1):52–61.

[Atlas et al., 1990] Atlas, L., Cohn, D., Ladner, R., El-Sharkawi, M., Marks II, R., Aggoune, M., and Park, D. (1990). Training connectionist networks with queries and selective sampling. *In Advances in Neural Information Processing Systems (NIPS)*, 3:566573.

[Aubrey et al., 2014] Aubrey, A., Christensen, L., Brockers, R., and Thompson, D. (2014). Greenhouse gas sensing using small unmanned aerial systems-field experiment results and future directions. *AGU Fall Meeting Abstracts*, 1:05.

[Bartz-Beielstein et al., 2005] Bartz-Beielstein, T., Lasarczyk, C., and Preuss, M. (2005). Sequential parameter optimization. *In Proceedings CEC-05*.

[Bearden, 1967] Bearden, J. A. (1967). X-ray wavelengths. *Reviews of Modern Physics*, 39(1):78.

[Blei and Jordan, 2006] Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Brochu et al., 2010a] Brochu, E., Brochu, T., and de Freitas, N. (2010a). A bayesian interactive optimization approach to procedural animation design. *In Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*.

[Brochu et al., 2010b] Brochu, E., Cora, V. M., and De Freitas, N. (2010b). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

[Bultel et al., 2015] Bultel, B., Quantin, C., and Lozach, L. (2015). Description of cotcat (complement to crism analysis toolkit). *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):3039–3049.

[Burkart et al., 2014] Burkart, A., Cogliati, S., Schickling, A., and Rascher, U. (2014). A novel uav-based ultra-light weight spectrometer for field spectroscopy. *Sensors Journal, IEEE*, 14(1):62–67.

[Busby, 2009] Busby, D. (2009). Hierarchical adaptive experimental design for gaussian process emulators. *Reliability Engineering and System Safety*, 94(7):11831193.

[Cabrera-Bosquet et al., 2012] Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., and Luis Araus, J. (2012). High-throughput phenotyping and genomic selection: The frontiers of crop breeding convergef. *Journal of integrative plant biology*, 54(5):312–320.

[Calandra et al., 2014] Calandra, R., Seyfarth, A., Peters, J., and Deisenroth, M. P. (2014). An experimental comparison of bayesian optimization for bipedal locomotion. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1951–1958. IEEE.

[Candela and Rasmussen, 2005] Candela, J. Q. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*.

[Caselton et al., 1992] Caselton, W. F., Kan, L., and Zidek, J. V. (1992). *Statistics in the Environmental and Earth Sciences*. Halsted Press.

[Castaño et al., 2005] Castaño, R., Anderson, R. C., Estlin, T., DeCoste, D., Fisher, F., Gains, D., Mazzoni, D., and Judd, M. (2005). Rover traverse science for increasedmission science return. *IEEE Aerospace*.

[Chaloner and Verdinelli, 1995] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3):273304.

[Cohn, 1994] Cohn, D. A. (1994). Neural network exploration using optimal experiment design. *Neural Networks*.

[Cornford et al., 2002] Cornford, D., Nabney, I. T., and Williams, C. K. (2002). Modelling frontal discontinuities in wind fields. *Journal of nonparametric statistics*, 14(1-2):43–58.

[Cox and John, 1992] Cox, D. D. and John, S. (1992). A statistical method for global optimization. *In Proc. IEEE Conference on Systems, Man and Cybernetics*, 2:12411246.

[Cox and John, 1997] Cox, D. D. and John, S. (1997). Sdo: A statistical method for global optimization. *Multidisciplinary Design Optimization: State of the Art*, page 315329.

[Cressie, 1991] Cressie, N. A. (1991). *Statistics for Spatial Data*. Wiley and Sons, New York.

[Dasgupta et al., 2008] Dasgupta, S., Hsu, D., and Monteleoni, C. (2008). A general agnostic active learning algorithm. *In Advances in Neural Information Processing Systems (NIPS)*, 20:353360.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B.*, 39 (1):138.

[Diaz et al., 2015] Diaz, J., Pieri, D., Wright, K., Sorensen, P., Kline-Shoder, R., Arkin, C., Fladeland, M., Bland, G., Buongiorno, M., Ramirez, C., Corrales, E., Alan, A., Alegria, O., Diaz, D., and Linick, J. (2015). Unmanned aerial mass spectrometer systems for in-situ volcanic plume analysis. *Journal of The American Society for Mass Spectrometry*, 26(2):292–304.

[Duin, 1976] Duin, R. (1976). On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Trans. Computers*, page 11751179.

[Escobar and West, 1995] Escobar, M. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*.

[Estlin et al., 2012] Estlin, T. a., Bornstein, B. J., Gaines, D. M., Anderson, R. C., Thompson, D. R., Burl, M., Castaño, R., and Judd, M. (2012). AEGIS Automated Science Targeting for the MER Opportunity Rover. *ACM Transactions on Intelligent Systems and Technology*, 3:1–19.

[Ferguson, 1973] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

[Foil et al., 2016] Foil, G., Fong, T., Deans, M., Elphic, R. C., and Wettergreen, D. (2016). Physical Process Models for Improved Rover Mapping. *iSAIRAS*.

[Foil et al., 2013] Foil, G., Thompson, D. R., Abbey, W., and Wettergreen, D. S. (2013). Probabilistic surface classification for rover instrument targeting. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 775–782.

[Fox et al., 2007] Fox, E. B., Sudderth, E. B., and Willsky, A. S. (2007). Hierarchical dirichlet processes for tracking maneuvering targets. In *Information Fusion, 2007 10th International Conference on*, pages 1–8. IEEE.

[Friedman et al., 2011] Friedman, A., Steinberg, D., Pizarro, O., and Williams, S. B. (2011). Active learning using a variational dirichlet process model for pre-clustering and classification of underwater stereo imagery. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1533–1539. IEEE.

[Gao et al., 2009] Gao, B. C., Montes, M. J., Davis, C. O., and Goetz, A. F. (2009). Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. *Remote Sensing of Environment*, 113:S17S24.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *PAMI*, page 721741.

[Goldberger and Roweis, 2005] Goldberger, J. and Roweis, S. (2005). Hierarchical clustering of a mixture model. *Advances in Neural Information Processing Systems*, page 505512.

[Gonzales, 1985] Gonzales, T. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, page 38(23):293306.

[Green et al., 1998] Green, R. O., Eastwood, M. L., Sarture, C. M., Chrien, T. G., Aronsson, M., Chippendale, B. J., Faust, J. A., Pavri, B. E., Chovit, C. J., Solis, M., and et al. (1998). Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris). *Remote Sensing of Environment*, 65(3):227248.

[Griffiths and Ghahramani, 2005] Griffiths, T. L. and Ghahramani, Z. (2005). Infinite latent feature models and the indian buffet process. *Gatsby Computational Neuroscience Unit, Tech. Rep. 2005-001*.

[Haines and Tao Xiang, 2013] Haines, T. S. F. and Tao Xiang (2013). Active Learning using Dirichlet Processes for Rare Class Discovery and Classication. *IJCV 2013*.

[Hamlin et al., 2011] Hamlin, L., Green, R., Mouroulis, P., Eastwood, M., Wilson, D., Dudik, M., , and Paine, C. (2011). Imaging spectrometer science measurements for terrestrial ecology: Aviris and new developments. *IEEE Aerospace Conference*.

[Huang et al., 2006] Huang, D., Allen, T. T., Notz, W. I., and Zheng, N. (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, page 34(3):441466.

[Hutter et al., 2009] Hutter, F., Hoos, H. H., Leyton-Brown, K., and Murphy, K. P. (2009). An experimental investigation of model-based parameter optimisation: Spo and beyond. *In Proceedings of GECCO09*.

[Hwangbo et al., 2009] Hwangbo, J. W., Di, K., Li, R., and Science, G. (2009). Integration of Oribital and Ground Image Networks for the Automation of Rover Localization. *ASPRS*.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

[Jones, 2001] Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, page 21:345383.

[Jones et al., 1998] Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal Global Optimization*, page 13(4):455492.

[Kang et al., 2004] Kang, J., Ryu, K., and Kwon, H. (2004). Using cluster-based sampling to select initial training set for active learning in text classification. *In Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, page 384388.

[Kapoor et al., 2007] Kapoor, A., Grauman, K., Urtasun, R., and Sarrell, T. (2007). Active learning with gaussian processes for object categorization. *In Proc. International Conference on Computer Vision (ICCV)*.

[Keinosuke and Hostetler, 1975] Keinosuke, F. and Hostetler, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, page 3240.

[Keshava and Mustard, 2002] Keshava, N. and Mustard, J. F. (2002). Spectral Unmixing. *IEEE Signal Processing Magazine*.

[Kohanbash et al., 2011] Kohanbash, D., Valada, A., and Kantor, G. A. (2011). Wireless sensor networks and actionable modeling for intelligent irrigation. In *2011 ASABE Annual International Meeting*.

[Krause et al., 2006] Krause, a., Guestrin, C., Gupta, a., and Kleinberg, J. (2006). Near-optimal sensor placements: maximizing information while minimizing communication cost. *5th International Conference on Information Processing in Sensor Networks*, pages 2–10.

[Krause et al., 2008] Krause, A., Singh, A., and Guestrin, C. (2008). Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284.

[Krause and Oliver, 1979] Krause, M. O. and Oliver, J. (1979). Natural widths of atomic k and l levels, k x-ray lines and several kll auger lines. *Journal of Physical and Chemical Reference Data*, 8(2):329338.

[Krishnan et al., 2013] Krishnan, A. K., McGarey, P., Saripalli, S., and Bell, J. F. (2013). Nir-cam : Development of a near infrared camera. *IEEE International Symposium on Robotic and Sensors Environments (ROSE)*.

[Kushner and Yin, 1997] Kushner, H. J. and Yin, G. G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer-Verlag.

[Lewis and Catlett, 1994] Lewis, D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. *In Proceedings of the International Conference on Machine Learning (ICML)*, page 148156.

[Lewis and Gale, 1994] Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc.

[Lizotte, 2008] Lizotte, D. (2008). Practical bayesian optimization. *PhD thesis, University of Alberta, Edmonton, Alberta, Canada*.

[Lloyd, 1982] Lloyd, S. P. (1982). Least squares quantization. *IEEE Transactions on Information Theory*, 28(2):129137.

[MartinezCantin et al., 2009] MartinezCantin, R., de Freitas, N., Brochu, E., Castellanos, J., and Doucet, A. (2009). A bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonomous Robots*, page 27(2):93103.

[McLachlan and Basford, 1988] McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.

[Mishkin et al., 2006] Mishkin, A. H., Limonadi, D., Laubach, S. L., and Bass, D. S. (2006). Working the martian night shift-the mer surface operations process. *IEEE robotics & automation magazine*, 13(2):46–53.

[Mockus, 1994] Mockus, J. (1994). Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, page 4(4):347 365.

[Mockus et al., 1978] Mockus, J., Tiesis, V., and Zilinskas, A. (1978). *Toward Global Optimization*, volume 2. Elsevier.

[Murchie et al., 2007] Murchie, S., Arvidson, R., Bedini, P., Beisser, K., Bibring, J.-P., Bishop, J., Boldt, J., Cavender, P., Choo, T., Clancy, R., et al. (2007). Compact reconnaissance imaging spectrometer for mars (crism) on mars reconnaissance orbiter (mro). *Journal of Geophysical Research: Planets*, 112(E5).

[NASA, 2016a] NASA (2016a). Adaptive caching concept - mars 2020 rover. `http://mars.nasa.gov/mars2020/multimedia/images/?ImageID=7291`. Accessed: 2016-09-30.

[NASA, 2016b] NASA (2016b). Instruments - mars 2020 rover. `http://mars.nasa.gov/mars2020/mission/instruments/`. Accessed: 2016-10-20.

[NASA, 2016c] NASA (2016c). Msl science corner - science operations. `https://msl-scicorner.jpl.nasa.gov/scienceoperations/`. Accessed: 2016-09-30.

[NASA, 2016d] NASA (2016d). Objectives - mars 2020 rover. `http://mars.nasa.gov/mars2020/mission/science/objectives/`. Accessed: 2016-10-20.

[NASA, 2016e] NASA (2016e). Planetary instrument for x-ray lithochemistry (pixl). `http://mars.nasa.gov/mars2020/mission/instruments/pixl/for-scientists/`. Accessed: 2016-09-30.

[Neal, 2007] Neal, R. M. (2007). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

[Nguyen and Smeulders, 2004] Nguyen, H. and Smeulders, A. (2004). Active learning using pre-clustering. *In Proceedings of the International Conference on Machine Learning (ICML)*, page 7986.

[Nuske et al., 2011] Nuske, S. T., Achar, S., Bates, T., Narasimhan, S. G., and Singh, S. (2011). Yield estimation in vineyards by visual grape detection. In *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '11)*.

[Pedersen et al., 2001] Pedersen, L., Wagner, M., Apostolopoulos, D., and Whittaker, W. (2001). Autonomous robotic meteorite identification in antarctica. *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation*, 4:4158–4165.

[Pingali, 2012] Pingali, P. L. (2012). Green revolution: Impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences*, 109(31):12302–12308.

[Poland, 2014] Poland, J. (2014). Genomics assisted breeding and field-based high throughput phenotyping. *Presented at DOE ARPA-E Advanced Plant Phenotyping Workshop, Chicago, IL*.

[Rasmussen, 1999] Rasmussen, C. E. (1999). The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560.

[Rasmussen, 2006] Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*. Citeseer.

[Ray et al., 2012] Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., and Foley, J. A. (2012). Recent patterns of crop yield growth and stagnation. *Nature communications*, 3:1293.

[Rosenblatt, 1956] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 (3):832.

[Roy and McCallum, 2001] Roy, N. and McCallum, A. (2001). Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*.

[Ryan et al., 1988] Ryan, C., Clayton, E., Griffin, W., Sie, S., and Cousens, D. (1988). Snip, a statistics-sensitive background treatment for the quantitative analysis of pixe spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 34(3):396402.

[Sasena, 2002] Sasena, M. J. (2002). Flexibility and efficiency enhancement for constrained global design optimization with kriging approximations. *PhD thesis, University of Michigan*.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Sebastiani and Wynn, 2000] Sebastiani, P. and Wynn, H. P. (2000). Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society*.

[Settles, 2009] Settles, B. (2009). Active learning literature survey. *Technical Report 1648, Uni. of Wisconsin-Madison*.

[Settles, 2010] Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

[Settles et al., 2008a] Settles, B., Craven, M., and Friedland, L. (2008a). Active learning with real annotation costs. *In Proceedings of the NIPSWorkshop on Cost-Sensitive Learning*.

[Settles et al., 2008b] Settles, B., Craven, M., and Ray, S. (2008b). Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296.

[Shewry and Wynn, 1987] Shewry, M. and Wynn, H. (1987). Maximum entropy sampling. *Journal of Applied Statistics*, 14(2).

[Sillito and Fisher, 2007] Sillito, R. and Fisher, R. (2007). Incremental one-class learning with bounded computational complexity. *Artificial Neural NetworksICANN 2007*, pages 58–67.

[Smith et al., 2011] Smith, B., Settles, B., Hallows, W., Craven, M., and Denu, J. (2011). Sirt3 substrate specificity determined by peptide arrays and machine learning. *ACS Chemical Biology*, 6(2):146157.

[Smith et al., 2005] Smith, T., Niekum, S., Thompson, D. R., and Wettergreen, D. (2005). Concepts for science autonomy during robotic traverse and survey. *2005 IEEE Aerospace Conference*.

[Snelson and Ghahramani, 2005] Snelson, E. and Ghahramani, Z. (2005). Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264.

[Snelson, 2007] Snelson, E. L. (2007). *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, University College London.

[Srinivas et al., 2010] Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. *In Proceedings of the International Conference on Machine Learning (ICML)*.

[Stein, 1999] Stein, M. L. (1999). Interpolation of spatial data: Some theory for kriging. *Springer Series in Statistics.*

[Streltsov and Vakili, 1999] Streltsov, S. and Vakili, P. (1999). A non-myopic utility function for statistical global optimization algorithms. *Journal of Global Optimization*, page 14:283298.

[Sudderth et al., 2005] Sudderth, E. B., Torralba, A., Freeman, W. T., and Willsky, A. S. (2005). Describing Visual Scenes using Transformed Dirichlet Processes. *Neural Information Processing Systems (NIPS)*.

[Swayze et al., 2014] Swayze, G. A., Clark, R. N., Goetz, A. F., Livo, K. E., Breit, G. N., Kruse, F. A., Sutley, S. J., Snee, L. W., Lowers, H. A., Post, J. L., et al. (2014). Mapping advanced argillic alteration at cuprite, nevada, using imaging spectroscopy. *Economic Geology*, 109(5):1179–1221.

[Teh et al., 2006] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

[Tesch et al., 2013a] Tesch, M., Schneider, J., and Choset, H. (2013a). Learning stochastic binary tasks using bayesian optimization with shared task knowledge. *International Conference on Machine Learning: Workshop on Robot Learning*.

[Tesch et al., 2013b] Tesch, M., Schneider, J. G., and Choset, H. (2013b). Expensive function optimization with stochastic binary outcomes. In *ICML (3)*, pages 1283–1291.

[Thompson, 2008] Thompson, D. R. (2008). *Intelligent Mapping for Autonomous Robotic Survey*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

[Thompson et al., 2014] Thompson, D. R., Allwood, A., Assad, C., Flannery, D., Hodyss, R., Knowles, E., and Wade, L. (2014). Adaptive sampling for rover x-ray lithochemistry. *International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)*.

[Thompson et al., 2015a] Thompson, D. R., Flannery, D. T., Lanka, R., Allwood, A. C., Bue, B. D., Clark, B. C., Elam, W. T., Estlin, T. A., Hodyss, R. P., Hurowitz, J. A., Liu, Y., and Wade, L. A. (2015a). Automating x-ray fluorescence analysis for rapid astrobiology surveys. *Astrobiology*, 15:961–976.

[Thompson et al., 2015b] Thompson, D. R., Foil, G., Furlong, M., Kiran, A. R., and Wettergreen, D. (2015b). Spatio-spectral exploration combining in situ and remote measurements. *AAAI*.

[Thompson et al., 2015c] Thompson, D. R., Gao, B.-C., Green, R. O., Roberts, D. A., Dennison, P. E., , and Lundeen, S. R. (2015c). Atmospheric correction for global mapping spectroscopy: Atrem advances for the hyspiri preparatory campaign. *Remote Sensing of Environment*, 167:6477.

[Thompson and Wettergreen, 2008] Thompson, D. R. and Wettergreen, D. (2008). Intelligent maps for autonomous kilometer-scale science survey. In *International Symposium on Artificial Intelligence, Robotics and Automation in Space (iSAIRAS)*.

[Tilman, 2011] Tilman, D. e. a. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, page 2026020264.

[Torn and Zilinskas, 1989] Torn, A. and Zilinskas, A. (1989). *Global Optimization*. Springer-Verlag.

[Van Grieken and Markowicz, 2001] Van Grieken, R. and Markowicz, A. (2001). *Handbook of X-ray Spectrometry*. CRC Press.

[Vatturi and Wong, 2009] Vatturi, P. and Wong, W.-K. (2009). Category detection using hierarchical mean shift. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–856. ACM.

[Wagstaff et al., 2013] Wagstaff, K. L., Thompson, D. R., Abbey, W., Allwood, A., Bekker, D. L., Cabrol, N. A., Fuchs, T., and Ortega, K. (2013). Smart, texture-sensitive instrument classification for in situ rock and layer analysis. *Geophysical Research Letters*, 40.

[Wendland, 2009] Wendland, H. (2009). *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press.

[Wettergreen et al., 2013] Wettergreen, D., Foil, G., Furlong, M., and Thompson, D. R. (2013). Science autonomy for rover subsurface exploration of the atacama desert. *AI Magazine*.

[Williams and Notz, 2000] Williams, B. J. andSantner, T. J. and Notz, W. I. (2000). Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, 10:11331152.

[Wood et al., 2006] Wood, F., Goldwater, S., and Black, M. J. (2006). A non-parametric Bayesian approach to spike sorting. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1:1165–8.

[Zhu and Goldberg, 2009] Zhu, X. and Goldberg, A. (2009). *Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan and Claypool.