# Robust Non-Gaussian Semantic Simultaneous Localization and Mapping

by

Kevin J. Doherty

B.E., Stevens Institute of Technology (2017)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science in Aeronautics and Astronautics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY AND

WOODS HOLE OCEANOGRAPHIC INSTITUTION

September 2019



Author .....
Department of Aeronautics and Astronautics
August 22, 2019

Certified by .....
John J. Leonard
Samuel C. Collins Professor of Mechanical and Ocean Engineering
Thesis Supervisor

Accepted by .....
Sertac Karaman
Associate Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Accepted by .....
David Ralston
Associate Scientist with Tenure, Applied Ocean Physics and Engineering
Chair, Joint Committee for Applied Ocean Science and Engineering

# Robust Non-Gaussian Semantic Simultaneous Localization and Mapping

by

Kevin J. Doherty

Submitted to the Department of Aeronautics and Astronautics
on August 22, 2019, in partial fulfillment of the
requirements for the degree of
Master of Science in Aeronautics and Astronautics

## Abstract

The recent success of object detection systems motivates *object-based* representations for robot navigation; i.e. semantic simultaneous localization and mapping (SLAM), in which we aim to jointly estimate the pose of the robot over time as well as the location and semantic *class* of observed objects. A solution to the semantic SLAM problem necessarily addresses the continuous inference problems *where am I?* and *where are the objects?*, but also the discrete inference problem *what are the objects?*.

We consider the problem of semantic SLAM under non-Gaussian uncertainty. The most prominent case in which this arises is from *data association uncertainty*, where we do not know with certainty what objects in the environment caused the measurement made by our sensor. The semantic class of an object can help to inform data association; a detection classified as a door is unlikely to be associated to a chair object. However, detectors are imperfect, and incorrect classification of objects can be detrimental to data association. While previous approaches seek to eliminate such measurements, we instead model the robot and landmark state uncertainty induced by data association in the hopes that new measurements may disambiguate state estimates, and that we may provide representations useful for developing decision-making strategies where a robot can take actions to mitigate multimodal uncertainty.

The key insight we leverage is that the semantic SLAM problem with unknown data association can be reframed as a non-Gaussian inference problem. We present two solutions to the resulting problem: we first assume Gaussian measurement models, and non-Gaussianity only due to data association uncertainty. We then relax this assumption and provide a method that can cope with arbitrary non-Gaussian measurement models. We show quantitatively on both simulated and real data that both proposed methods have robustness advantages as compared to traditional solutions when data associations are uncertain.

Thesis Supervisor: John J. Leonard
Title: Samuel C. Collins Professor of Mechanical and Ocean Engineering

# Acknowledgments

I would first and foremost like to thank Professor John Leonard for giving me the opportunity to join the Marine Robotics Group at MIT and being a constant source of inspiration and positivity in robotics. His advice, encouragement, and support made the work in this thesis possible. It is because of him that I first learned of the MIT-WHOI Joint Program, for which I am incredibly grateful.

I would also like to thank the members of the Marine Robotics Group and Robust Robotics Group, especially Prof. Nicholas Roy for giving me a "home base" at MIT in the Robust Robotics Group during my first year. It's hard to imagine a better group of folks to chat with about research, and I appreciate so much the stimulating research discussion, positivity, and support within these groups. I feel so lucky to have had the opportunity to learn from them.

I'd like to thank my MIT-WHOI Joint Program cohort for being great friends and really making Cambridge and Woods Hole my home for the past two years. I'd also like to acknowledge the program more broadly for giving me the opportunity to work on exciting problems and meet such incredibly talented and multi-faceted people.

Finally, I would like to thank my mom, my dad, Matt, Karyn, and Ashley; their love and support over the years has truly been invaluable.

# Funding

This work was partially supported by the Office of Naval Research under grants N00014-18-1-2832 and N00014-16-2628, as well as the National Science Foundation (NSF) Graduate Research Fellowship.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis concerns the problem of *data association* for object-level navigation. Data association is the process of determining the correspondence between a sensor measurement and a landmark. Humans very reliably perform data association. Without conscious effort we are aware of the correspondence between what we see and what exists in our environment. It is often these tasks, which even a child can perform reliably, that are most difficult for machines.

Humans navigate with respect to both *geometric* cues, avoiding collisions with the environment, as well as *semantic* cues, bearing in mind the history of objects observed [15]. For example, taking a walk around a city block in an unfamiliar area, we may determine that we've come full circle through the recognition of cars and houses that we had previously observed. The embodiment of these capabilities into analogous robot systems can permit similarly reliable navigation and reasoning over the large spatial and temporal scales that humans are accustomed to. Scaling navigation in this sense is a more critical challenge now than ever before as autonomous robots are deployed in increasingly challenging situations.

We consider the problem of determining the associations of measurements to landmarks for robot navigation. Specifically we deal with the use of objects as environmental landmarks. Formally, we will take an object to be defined uniquely by its position in space and semantic class, where the class of an object is a label from a predefined set of types of objects we can recognize. As an example, we may have a mo-

bile robot equipped with a perceptual system capable of detecting and classifying cars (like those detected in Figure 1-1), houses, and pedestrians. The objective of the data association system, then, is to use the perceived geometric and semantic information in conjunction with estimates of the location and classes of environmental landmarks to place the measurements in correspondence with the landmarks. The challenge for such a system is that while in principle the geometric and semantic information aids in associating measurements to known landmarks, errors in object classification from a detector, as well as errors in vehicle and landmark position estimation cause challenges for establishing reliable measurement-landmark correspondence.

Until recently, poor classification performance made the use of object detections impractical, and most state-of-the-art visual SLAM algorithms (e.g. [39], [16]) neglect semantics in favor of tracking sparse, relatively uninterpretable visual descriptors or measuring photoconsistency between image frames as a means of determining a camera's relative pose change between instances when the images were obtained. However, the recent performance of object detectors like the "single-shot multibox detector" (SSD) [35] and "You Only Look Once" (YOLO) [46], among many others, has motivated further the use of semantics to provide a sparse set of interpretable visual cues for mobile robot navigation. Moreover, semantic information may go beyond class alone, and can involve reasoning about the uses, dynamics, and deformabiity of objects. This places the problem of semantic SLAM in a position to unify mobile robot navigation and scene understanding, a task which we hope will bring long-term robot navigation performance to the level of reliability of humans and beyond.

## 1.1 Motivation

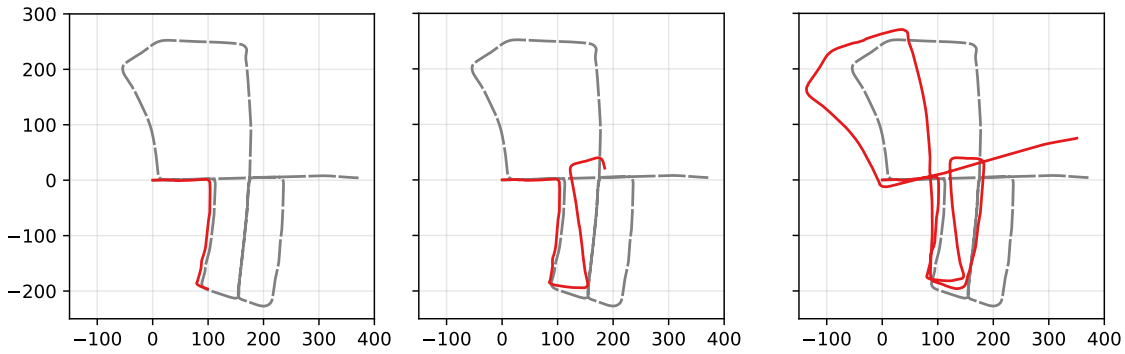
Robots are no longer isolated to factories and manufacturing applications. In recent years, we have seen a number of applications of robots operating in complex environments and in safety-critical situations. The critical navigation challenges for these platforms have not changed much in the past 30 years, but our notion of success in these tasks has changed. Modern robots are expected not just to move from one point

Figure 1-1: Detection of a car produced by a neural network using imagery from the KITTI dataset [22]. The detector fails to detect all of the cars in the scene and moreover produces an ambiguous bounding box containing several cars. We aim to develop object-level navigation systems that are robust to these types of detections.

to another without collision, but to do so with low-cost sensors, reason about their uncertainty to do so while incurring minimal localization error, and deduce objectives from high-level task specifications (requiring building and reasoning about a semantic world representation).

Moreover, as object detectors continue to improve, there has been growing interest in their use in conjunction with a suite of inertial and geometric sensors to improve robot navigation, allowing autonomous robots to build more accurate, descriptive maps [6, 53]. However, the consideration of semantic class and data association ambiguity makes the navigation problem significantly more challenging. The majority of SLAM systems are well-suited to solving a specific subset of navigation problems in which sensor measurements can be represented as nonlinear functions of the true state corrupted by additive Gaussian noise (see, for example, [6] for a review of state-of-the art methods making these assumptions). When we consider jointly the discrete inference problem and the continuous inference problem (even under the same measurement assumptions), ambiguity in the discrete variables introduces *multiple hypotheses* about the robot and world state. The resulting "multi-hypothesis" ambiguity is generally non-Gaussian and is poorly addressed by traditional methods for SLAM.

In particular, in Figure 1-2 we show three snapshots of a trajectory estimate for a vehicle equipped with stereo cameras (from the popular KITTI dataset [22]). De-




(a) 16% poses added. (b) 33% poses added. (c) 100% poses added.

Figure 1-2: Snapshots of the landmark-based SLAM solution using maximum-likelihood data association **(red)** for the KITTI dataset odometry sequence 05 with ground-truth **(gray, dashed)**. Detections of cars (see Figure 1-1) are used as landmarks with range and bearing provided by stereo cameras. Initially, in 1-2a the solution is consistent with the ground-truth, but as incorrect loop closures are added in 1-2b and beyond in 1-2c the trajectory estimate fails catastrophically.

tections of cars, like those in Figure 1-1 are used as landmarks and their position in space is determined by the average range and bearing to the three-dimensional points recovered by the stereo cameras falling in the detection bounding box. The methods producing the pictured trajectories make traditional assumptions of Gaussian measurement uncertainty and use maximum-likelihood data association. The maximum-likelihood data association method determines, for each observation, the most likely landmark to have caused the observation (a detailed description of the maximum-likelihood method for data association is given in Chapter 3). We find that the method aligns well with the ground-truth initially, but errors arise when the most likely association at a particular time is not the true association. We refer to these events as false loop closures. One example of a false loop closure can be seen in Figure 1-2b, and we see how catastrophically incorrect the overall trajectory estimate (shown in Figure 1-2c) can be as these errors propagate to cause errors in future associations.

Beyond recent research and commercial interest in autonomous driving vehicles and quadrotor drones, one of the most compelling areas for autonomy—and one of the most challenging—is marine robotics. However, all of the aforementioned problems are made even more challenging by the underwater environment: underwater scenes are often repetitive, ambiguous, or featureless when they can be observed with

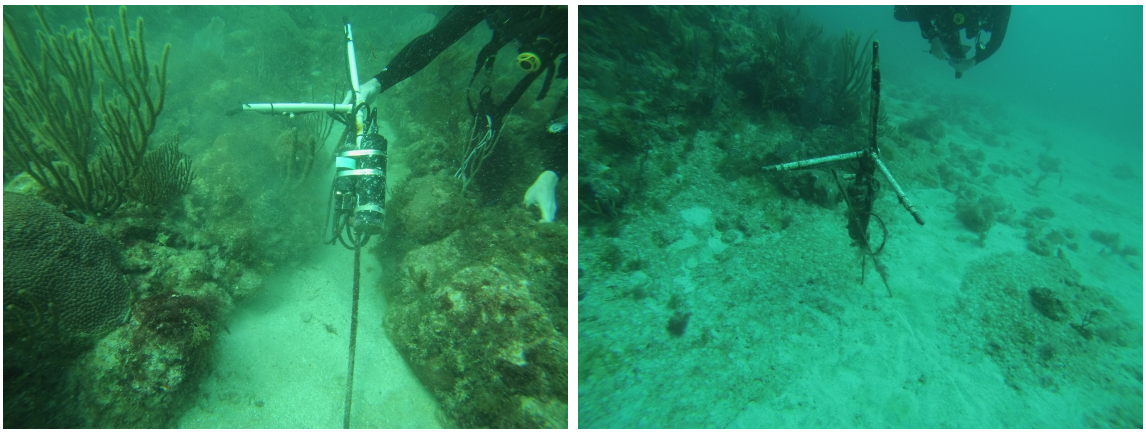


Figure 1-3: Instrument recovery for underwater science is a compelling area for autonomous vehicles. The soundtraps in these photos, located off the coast of the United States Virgin Islands were displaced during the back-to-back category 5 hurricanes Irma and Maria. Recovery required a team of several divers and a total dive time of 771 minutes. Images courtesy of Genevieve Flaspohler.

cameras, and acoustic sensing, one of the primary modes of perception for underwater vehicles, suffers greatly from non-Gaussian noise sources like acoustic multipath. Because of the difficulties of operating in the underwater environment, a number of mundane or dangerous tasks have historically been performed by human divers. For example, the recovery of soundtraps for biological research in Figure 1-3, and the dangerous task of inspecting ship hulls for mines (which has previously been performed by either trained human divers or marine mammals). Additionally, there has been increasing interest in the use of robots for semantic-level biological surveying, as in prior work which examines online learning for multiple communication-constrained underwater vehicles [12].

Moreover, there is increasing interest in the development of low-cost underwater drone platforms, like the BlueRobotics BlueROV2 [1]. In order for these vehicles to be practically useful for many of the most prominent underwater tasks, there is a critical need for robust perception and state estimation algorithms that can cope with the difficulties of operating in underwater environments.

These challenges for the next generation of autonomous systems motivate richer representations of uncertainty that can consider semantics and sensor noise models with complex noise distributions. Such representations accommodate the errors

in data association caused by geometric uncertainty in bounding boxes provided by object detectors (such as in Figure 1-1), as well as the environment dependent, non-Gaussian noise incurred by a variety of sensors (like sonars in underwater environments).

## 1.2 Related Work

There is a rich history of literature on the problems of SLAM and data association. Early work on probabilistic data association (PDA) as a representation for ambiguous hypotheses stems from target-tracking literature, where it was incorporated into the "probabilistic data association filter" [3]. Approaches based on multi-hypothesis tracking (MHT) originated around the same time [47], later adapted to the SLAM problem [8], [9]. These approaches seek to explicitly represent several plausible hypotheses and over time "prune" those which become unlikely.

In the recent history of SLAM, there have been three primary areas of related work: *robust* SLAM, *non-Gaussian* SLAM, and *semantic* SLAM. Each of these areas has substantial overlap with the others, but here we attempt to place each work into its most relevant subcategory.

### 1.2.1 Robust SLAM

A number of works in *robust SLAM* address the problem of SLAM with outliers. These can alternatively be viewed as methods that deal with discrete-continuous estimation, where the discrete variables are decisions about whether to discard particular measurements. Sunderhauf et al. [55] introduce discrete switching variables which are estimated on the back-end to determine whether loop closure proposals from the front-end are accepted. Relatedly, Latif et al. [33] introduce a method for "undoing" incorrect loop closures. Olson and Agarwal [42] proposed a "max-mixtures" approach that side-steps the complexity usually associated with multi-hypothesis SLAM or non-Gaussian SLAM by selecting the most likely component of a mixture of Gaussians at all points in the measurement domain (where each Gaussian corresponds to

a candidate hypothesis). This removes the need to explicitly represent a potentially large number of combinations of hypotheses. Pfingsthorn and Birk [44] proposed a maximum-likelihood optimization for multimodal distributions. In [49], the authors propose a method for approximate inference on a junction tree data structure for robust SLAM problems with Gaussian measurement models.

Similar to [42], we consider a max-mixture-style approach to inference of the most probable set of robot poses and landmark locations in Chapter 4. In contrast to their work, however, we consider mixtures of associations between different landmarks to address the data association problem, whereas they provide a method of rejecting incorrect loop closures. This difference motivates a new method for computing mixture component weights, whereas mixture weights are determined heuristically in [42]. Furthermore, this method, as well as the method we propose in 4 are suited to problems where individual measurements are well-characterized by Gaussian distributions but non-Gaussianity arises due to the introduction of discrete variables only. In Chapter 5 we expand further on these previous solutions by relaxing these assumptions.

### 1.2.2 Non-Gaussian SLAM

We divide the area of SLAM with non-Gaussian noise models into two categories: 1) methods that cope with arbitrary non-Gaussian distributions, including non-Gaussianity that arises from nonlinearity, undetermined systems (such as range-only or bearing-only SLAM), discrete variables (such as data association), or measurement physics (for example acoustic multipath in sonar measurements), and 2) methods that deal only with discrete-continuous Gaussian models.

#### Methods for Arbitrary Distributions

In the SLAM literature, FastSLAM [37] represents multiple hypotheses using a particle filter-based algorithm in which data association probabilities are computed separately for each particle representing a candidate robot state. Conceptually, FastSLAM is similar to our approach, but maintains separate parametric solutions each using

extended Kalman filters (EKFs). In contrast, our approach directly approximates the non-Gaussian solution to the SLAM problem under ambiguous association. A similar approach focusing on filtering-based SLAM is the *sum of Gaussians* method described by Durrant-Whyte et al. [14]. These methods infer the belief in the robot pose at the most recent discrete time given the history of measurements. We consider in this thesis methods which use all measurements to inform inference of the robot pose at all times in the history of the vehicle's trajectory (i.e. including the influence of measurements that took place *after* a particular time). In this way, using the methods presented in Chapter 5 we can recover a complex, non-Gaussian belief over the entire history of vehicle locations.

**Methods for Discrete-Continuous Gaussian Models**

When non-Gaussianity exists only due to the introduction of discrete association variables, the SLAM problem is often reduced to a search over discrete variables combined with a continuous optimization over vehicle poses and landmarks. *Multi-hypothesis* SLAM methods maintain tree data structures that model the sequence of discrete data association decisions. Each set of decisions has a corresponding solution to the continuous optimization problem. Examples of multi-hypothesis methods, besides the early works of [8] and [9], include [25], which described the concept of "lazy" data association. These methods generally seek to "prune" the most unlikely hypotheses to avoid the complexity of search over every possible assignment to discrete variables. Additionally, as previously mentioned, robust SLAM methods are similarly concerned with inference over discrete-continuous models, though historically the majority of robust SLAM methods have sought to deal with such discrete variables through optimization, with the aim of removing or coping with outliers, rather than multi-hypothesis search, where each hypothesis is in principle valid, but not necessarily correct.

### 1.2.3 Semantic SLAM

The ability of semantic measurements obtained from an object detector to aid in data association when there is ambiguity in purely geometric features links the problems of semantic SLAM and data association. The majority of works in semantic SLAM thus far have considered questions of geometric representation and make use of variants on maximum-likelihood data association [54, 48, 58, 36, 41]; in this thesis, we instead opt for a simple geometric representation and focus on representing the multimodalities induced in the posterior by ambiguous data associations and unknown landmark classes. Bowman *et al.* [5] recently showed that the discrete problems of landmark class inference and data association could be combined and provided an expectation-maximization (EM) solution which replaces the marginalization over data associations in the PDA method with a geometric mean, preserving the Gaussian assumption. The EM formulation provably converges to a local optimum when iterated, but for computational reasons, it is undesirable to recompute the combinatorial number of plausible data associations for previous poses. We also proactively compute data association probabilities, but marginalize out data associations (either by computing the "max marginal" or the true "sum marginal") and perform inference in the resulting factor graph, which is Gaussian in the former case, but non-Gaussian in the latter case.

## 1.3 Thesis Overview

In this thesis, we consider two complementary approaches to the non-Gaussian semantic SLAM problem. In one approach, we represent non-Gaussian noise as nonlinearity within a standard nonlinear least-squares SLAM framework. We show that this approach permits efficient inference and provides improved robustness to non-Gaussian noise as compared to current state-of-the-art alternatives. The second approach instead provides a representation of the full non-Gaussian posterior distribution over poses and landmarks. While this approach in its current form is less computationally efficient than the alternative, it is well-suited to provide the underlying representa-

tion for non-Gaussian belief space planning tasks. As a consequence, the latter work opens several new avenues for research connecting SLAM with active, non-Gaussian planning, for example reasoning about, and making decisions to mitigate data association uncertainty. Furthermore, we provide new theoretical insight linking these approaches as instances of max-product and sum-product belief propagation algorithms, respectively, for inference in graphical models.

# Chapter 2

# Perception as Bayesian Inference

In robot navigation, we are concerned with answering three questions: "where am I?". "where am I going", and "how should I get there?" [34]. However, robots perceive their environment only through noisy measurements. In order to navigate we must aggregate noisy measurements of the environment and robot state in the hopes of accurately determining the true location of the vehicle and an accurate map of the surroundings. We commonly describe the true state of the vehicle and environment as hidden, or *latent*, in that we can only infer them from measurements caused by the interaction of sensors with the environment. Since there is no way for us to know that the value of a particular inferred state matches the true state in general—if we knew with certainty the true state, we would be done—we typically express the inferred state as a probability distribution, termed the *belief* in the robot state. Mathematically, we formalize these ideas in the context of Bayesian inference, where belief is expressed as a posterior distribution over the latent variables–namely, the state of the robot and environment, described often as the position and orientation of the vehicle and relevant landmarks–conditioned on the *evidence*, given as the measurements obtained from sensors like lidar, cameras, sonar, or accelerometers.

In this chapter, we introduce robot perception as Bayesian inference. We begin by formally stating the SLAM problem as one of inference over a Bayesian network exposing probabilistic constraints between the sensor measurements and the hidden

states of the robot and landmarks [43]. We then discuss relevant background on inference in graphical models for navigation problems, including present solutions that are able to achieve reliable real-time navigation under specific assumptions about the robot's sensor noise characteristics. We will discuss the refactorization of graphical models into a tree structure (termed the Bayes tree in the context of robot navigation [29]). Finally, the principal algorithms for inference in graphical models, namely the *sum-product algorithm* and *max-product algorithm* for variable elimination will be discussed. These two algorithms form the cornerstone of this thesis. We will specifically show that the two core methods presented in this thesis can be derived as max-product and sum-product algorithms for sensors with sensor noise characteristics that can be described by mixtures of Gaussians.

## 2.1 Problem Formulation

As described in the previous section, during navigation a robot moves through the environment guided by a set of control inputs while making observations $\boldsymbol{z}_t \triangleq \{\boldsymbol{z}_t^1, \ldots, \boldsymbol{z}_t^K\}$ at each discrete time $t \in \mathbb{Z}_{\geq 0}$; here we use $K$ to denote the number of observations made at time $t$ (when this can vary for different points in time, we commonly denote the number of measurements made at a particular time $t$ as $K_t$). We will nominally consider all measurements to be real-valued vectors. The time index at the *end* of a robot-trajectory will generally be denoted as $T$.

In order to navigate, a robot must infer its state $\boldsymbol{x}_t$ in some state-space $\mathcal{X}$, and the state of all relevant environmental landmarks (i.e. the map), $\mathbf{L} \triangleq \{\boldsymbol{\ell}_j \in \mathcal{L},\ j = 1, \ldots, M\}$. The two tools we are equipped with to infer a belief over the latent robot and environment state are the *process* model and the *observation* or *measurement* model. The process model $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ describes how a robot's state may change stochastically from one discrete time instance to the next. This may represent, for example, constraints due to the dynamics of the vehicle, or the output of an inertial measurement device. The observation model $p(\boldsymbol{z}_t \mid \boldsymbol{x}_t, \boldsymbol{\ell}_j)$ describes the probability of making observation(s) $\boldsymbol{z}_t$ given that the robot state is $\boldsymbol{x}_t$ and knowledge of the

corresponding landmark state $\boldsymbol{\ell}_j$.

### 2.1.1 Problem Statement

In its general form, the problem of *semantic SLAM* is corresponds to determining the history of robot poses $\mathbf{X} \triangleq \{\boldsymbol{x}_t\}_{t=1}^T$ and landmark positions and semantic classes $\mathbf{L} \triangleq \{(\boldsymbol{\ell}^p, \boldsymbol{\ell}^s)_j\}_{j=1}^M$ given a set of sensor measurements $\mathbf{Z} \triangleq \{\boldsymbol{z}_t\}_{t=1}^T$ made at each robot pose. We have $\boldsymbol{x}_t \in SE(2)$ in the two-dimensional case, and $\boldsymbol{x}_t \in SE(3)$ in 3D. Similarly, we take $\boldsymbol{\ell}^p \in \mathbb{R}^2$ in 2D and $\mathbb{R}^3$ in 3D. The landmark class $\boldsymbol{\ell}^s$ is assumed to come from a finite set of discrete, known, class labels: $\mathcal{C} = \{1, 2, \ldots, C\}$.

More specifically, we aim to infer one or both of the following given a set of measurements $\mathbf{Z}$: (1) the posterior distribution over latent variables $\mathbf{X}$ and $\mathbf{L}$, conditioned on measurements $\mathbf{Z}$, or (2) the maximum *a posteriori* (MAP) estimates for $\mathbf{X}$ and $\mathbf{L}$. From the measurement and process models described in the previous section, we can write the SLAM problem in terms of a *Bayesian network*, or "Bayes net." A Bayes net is simply a directed, acyclic graph that encodes a joint probability distribution of the form:

$$p(\mathbf{V}) = \prod_i p(V_i \mid \mathbf{\Pi}_i), \tag{2.1}$$

where $V_i$ is a variable and $\mathbf{\Pi}_i$ is the set of "parents" of $V_i$. An example of a Bayesian network is given in Figure 2-1.

The posterior distribution over poses and landmarks i.e. $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$, can be written as being proportional to the joint distribution (assuming uniform priors on $\mathbf{X}$ and $\mathbf{L}$, excluding $\boldsymbol{x}_0$):

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto p(\boldsymbol{x}_0) \prod_{t=0}^{T} p(\boldsymbol{z}_t \mid \boldsymbol{x}_t, \boldsymbol{\ell}_{1:M}) \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}). \tag{2.2}$$

For example, the Bayes net in Figure 2-1 describes the joint distribution:

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{z}_1, \boldsymbol{z}_2) = p(\boldsymbol{z}_1 \mid \boldsymbol{x}_1, \boldsymbol{\ell}_1) p(\boldsymbol{z}_2 \mid \boldsymbol{x}_2, \boldsymbol{\ell}_2) p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) p(\boldsymbol{x}_1) p(\boldsymbol{\ell}_1) p(\boldsymbol{\ell}_2), \tag{2.3}$$

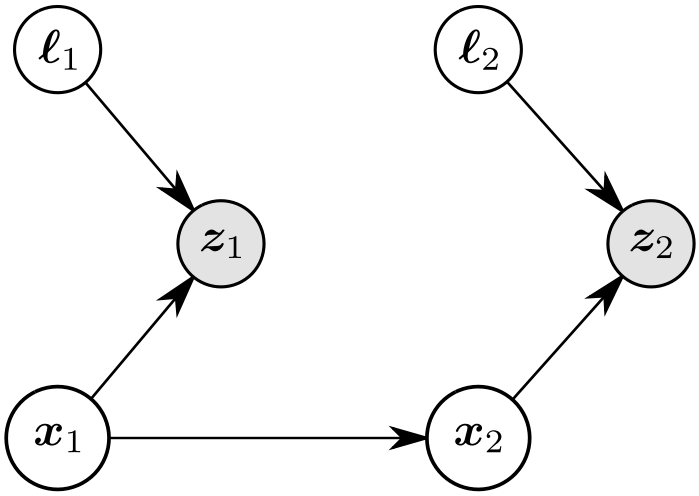



Figure 2-1: Bayesian network representation of simple SLAM problem. We depict latent variables, i.e. poses and landmarks, as unfilled circles and observed variables, in this case the sensor measurements, as filled circles.

which is, up to a constant of proportionality, equal to the posterior $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ for that example.

The MAP estimates for $\mathbf{X}$ and $\mathbf{L}$ can be obtained as:

$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}). \tag{2.4}$$

Maximization can be done after recovering the posterior, or as an alternative to full posterior inference, as we will describe later.

## 2.2 Probabilistic Modeling and Inference

### 2.2.1 Specifying the Process and Observation Models

Thus far, we have formulated the problem of robot navigation assuming the existence and knowledge of inter-pose measurement models (the *process* model for consecutive poses), and pose-landmark measurement models (the *observation* model). We will now clarify a few common probabilistic models for the types of measurements considered in this thesis.

We consider fairly simplistic perceptual noise models in this work, though the

framework we present is by no means limited to the models considered here. Specifically, we consider common nonlinear Gaussian noise models for all of our sensors, and show that even under these fairly stringent noise assumptions, there are challenging inference problems that can be posed within the context of semantic navigation. A more detailed treatment of nonlinear Gaussian measurement models of the form considered here can be found in [11], and a similar treatment for the case of non-Gaussian measurements is given in [19].

### Odometry Measurements

Odometry measurements may be obtained from a number of sources, including an inertial measurement unit (IMU), integration of wheel rotations, or visual feature tracking. In all of these systems the output is represented as the change in vehicle pose from a discrete time step $t-1$ to the subsequent time step $t$. We can represent the true change as a deterministic, nonlinear function of $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$. In particular, we have the function $h : \mathrm{SE}(d) \times \mathrm{SE}(d) \to \mathrm{SE}(d)$, defined as:

$$h(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) \triangleq \ominus\, \boldsymbol{x}_{t-1} \oplus \boldsymbol{x}_t. \tag{2.5}$$

Recall that poses are represented as elements of the special Euclidean group in either two or three dimensions. Here, we are making use of the pose inverse operation $\ominus$ and pose composition $\oplus$, which correspond to matrix inversion and matrix multiplication, respectively (see Appendix A.2 for a detailed description of operations on poses). The output of this nonlinear function, then, is also an element of special Euclidean group of the same dimensionality and represents the location and orientation of $\boldsymbol{x}_t$ in the reference frame of $\boldsymbol{x}_{t-1}$.

No odometry method, however, obtains the true transformation between two poses. Rather, the true transformation is corrupted by some noise in the measurement process. It is common to assume that the measurement is corrupted by additive Gaussian noise, and to properly account for the manifold structure of poses, it is common to assume that the measurement noise is Gaussian in the tangent space of

the measurement. In particular, let $\mathbf{T}$ denote the true transformation from $\boldsymbol{x}_{t-1}$ to pose $\boldsymbol{x}_t$ (i.e. $h(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)$), and $\tilde{\mathbf{T}}$ denote the measured transformation.

$$\tilde{\mathbf{T}} = \mathbf{T}\ \mathrm{Exp}(\epsilon),\ \epsilon \sim \mathcal{N}(0, \Sigma), \tag{2.6}$$

where Exp denotes the exponential map, taking an element of the tangent space to its counterpart in SE($d$). Background on manifolds for geometric representation in SLAM, including description of the exponential and logarithm maps, is provided in Appendix A. After some manipulation (see for example Forster et al. [18]) we can obtain that the distribution of the pose $\mathbf{T} = h(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)$ can be written in terms of the measurement $\tilde{\mathbf{T}}$ as:

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \eta(\tilde{\mathbf{T}}) \exp\left\{-\frac{1}{2}\|\mathrm{Log}(\mathbf{T}^{-1}\tilde{\mathbf{T}})\|^2_\Sigma\right\} \tag{2.7}$$

where $\eta(\tilde{\mathbf{T}})$ is a normalizing constant dependent only the measured transform $\tilde{\mathbf{T}}$ chosen to ensure the distribution integrates to 1, and Log denotes the logarithm map for SE($d$)[1]. The term $\|\cdot\|^2_\Sigma$ refers to the squared Mahalanobis distance, defined for a vector $\boldsymbol{v}$, given $\Sigma$ as:

$$\|\boldsymbol{v}|^2_\Sigma = \boldsymbol{v}^T\Sigma^{-1}\boldsymbol{v}. \tag{2.8}$$

**Range-Bearing Measurements**

Robots may also be equipped with sensors capable of measuring range and bearing, such as stereo cameras or lidar. A range sensor measures the Euclidean distance between the vehicle and a landmark:

$$r(\boldsymbol{x}, \boldsymbol{\ell}) \triangleq \|\boldsymbol{\ell}^p - \mathbf{t}\|_2, \tag{2.9}$$

[1]There has been substantial research on the representation of uncertainty on manifolds. The critical component that we make use of in this thesis is simply the resulting probabilistic model $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$. Consequently, we do not delve into much detail on the representation of spatial uncertainty or manifolds. For a thorough treatment of these topics, we refer the reader to [4].

where $\boldsymbol{\ell}^p$ is the position vector of the landmark and $\mathbf{t}$ is the translation vector of the pose $\boldsymbol{x}$. It is common to assume the range measurement $\boldsymbol{z}^r$ is equal to $r(\boldsymbol{x}, \boldsymbol{\ell})$ plus additive Gaussian noise, resulting in the sensor model:

$$p(\boldsymbol{z}^r \mid \boldsymbol{x}, \boldsymbol{\ell}) = \mathcal{N}(r(\boldsymbol{x}, \boldsymbol{\ell}); \mu_r, \sigma_r^2). \tag{2.10}$$

Bearing measurements similarly can be written as a nonlinear measurement. For example, the 2D bearing between a vehicle and a landmark can be written as:

$$\boldsymbol{\ell}^{\boldsymbol{x}} = \ominus\, \boldsymbol{x} \cdot \boldsymbol{\ell} \tag{2.11}$$

$$b(\boldsymbol{x}, \boldsymbol{\ell}) = \text{atan2}(\boldsymbol{\ell}_y^{\boldsymbol{x}}, \boldsymbol{\ell}_x^{\boldsymbol{x}}) \tag{2.12}$$

where here $\boldsymbol{\ell}^{\boldsymbol{x}}$ is the location of the landmark $\boldsymbol{\ell}$ expressed in the vehicle frame and $\cdot_x$ denotes the $x$-component of a vector (likewise for $y$). In 3D, we consider the bearing as well as elevation angle, $\text{atan2}(\boldsymbol{\ell}_z^{\boldsymbol{x}}, \boldsymbol{\ell}_x^{\boldsymbol{x}})$ and overload the $b(\boldsymbol{x}, \boldsymbol{\ell})$ notation to produce a vector containing bearing and elevation in the 3D case. As with range measurements, we typically assume additive Gaussian noise to obtain the probabilistic measurement model:

$$p(\boldsymbol{z}^b \mid \boldsymbol{x}, \boldsymbol{\ell}) = \mathcal{N}(b(\boldsymbol{x}, \boldsymbol{\ell}); \mu_b, \sigma_b^2), \tag{2.13}$$

where in 3D the $2 \times 2$ covariance matrix $\Sigma_b$ is used in place of $\sigma_b^2$.

### Semantic Measurements

Many modern object detectors are designed to produce a bounding box and class prediction for each object in an image frame. A bounding box can be used to inform the location of a landmark, but the class prediction provides semantic information that can aid in reliably associating new measurements to previously detected landmarks.

In order to use the semantic class prediction from an object detector, we make

use of the following probabilistic sensor model for class predictions:

$$p(\boldsymbol{z}^s \mid \boldsymbol{\ell}) = p(\boldsymbol{z}^s \mid \boldsymbol{\ell}^s) = \text{Cat}(CM(\boldsymbol{\ell}^s)) \tag{2.14}$$

where $\boldsymbol{z}^s \in \mathcal{C}$ is the semantic class prediction from the detector and $\boldsymbol{\ell}^s \in \mathcal{C}$ is the true semantic class of the landmark. Both are elements of a discrete set of *a priori* known classes $\mathcal{C}$. Here $CM(\boldsymbol{\ell}^s)$ denotes the normalized *confusion matrix* for the object detector indexed by the true class $\boldsymbol{\ell}^s$. For example, consider a general $C$ class classification problem, $\mathcal{C} = \{1, 2, \ldots, C\}$. A detector may have the following confusion matrix:

$$CM \quad = \quad \begin{array}{c} \\ \\ \text{Predicted class} \end{array} \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ C \end{array} \begin{array}{c} \text{True class} \\ \begin{array}{cccc} 1 & 2 & \ldots & C \end{array} \\ \begin{bmatrix} 0.9 & 0.05 & \ldots & 0.01 \\ 0.05 & 0.9 & \ldots & 0.01 \\ \vdots & \vdots & \ddots & \vdots \\ 0.01 & \ldots & \ldots & 0.9 \end{bmatrix} \end{array}. \tag{2.15}$$

In general, we assume that the object must have class in $\mathcal{C}$, so the columns of $CM$ must sum to 1. This is despite the fact that detectors may output detections even when there is no object with class in $\mathcal{C}$ in the bounding box. The probability $p(\boldsymbol{z}^s = i \mid \boldsymbol{\ell}^s = j)$ where $i, j \in \mathcal{C}$ corresponds to the $(i, j)$ element of the normalized confusion matrix. $\text{Cat}(\cdot)$ in the sensor model (2.14) denotes the categorical probability distribution parameterized by the given argument. In our case, the probability vector $CM(\boldsymbol{\ell}^s)$, the column of the confusion matrix corresponding to the true landmark class, parameterizes the categorical distribution.

The confusion matrix and consequently the probabilistic semantic measurement model can be obtained offline through validation on data. That is, given a dataset for which the true class of all objects is known, we can run a detector to obtain a set of predictions and compute the confusion matrix as the count of predicted classes for each object's true class. Normalizing the result gives the confusion matrix

above. Importantly, this process neglects the possibilities of missed detections or false positives (detections where there is no object in $\mathcal{C}$). Characterization of the object detector depends also on knowledge of the probabilities of these events, though missed detections are less problematic for navigation than false positives. Later in this thesis, we discuss the practical mitigation of false positives, but we do not explicitly model the probability of false positives or missed detections. A more detailed characterization of semantic sensor measurements that explicitly models these possibilities can be found in [2].

#### Combining Independent Measurements

Given the previously described sources of information, if we assume the measurement noise in each is independent (i.e. classification error does not depend on range or bearing, nor does range depend on class or bearing, and so on), we easily obtain the following factored measurement model for the combined measurement:

$$p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\ell}) = p(\boldsymbol{z}^r \mid \boldsymbol{x}, \boldsymbol{\ell}) p(\boldsymbol{z}^b \mid \boldsymbol{x}, \boldsymbol{\ell}) p(\boldsymbol{z}^s \mid \boldsymbol{\ell}). \tag{2.16}$$

The assumption of measurement independence is convenient for combining different measurement sources, but measurement errors may have nontrivial correlations. That said, if appropriate data is available offline, the full joint distribution for a measurement $\boldsymbol{z} \triangleq \{\boldsymbol{z}^r, \boldsymbol{z}^b, \boldsymbol{z}^s\}$ can be characterized empirically.

### 2.2.2 Graphical Models

#### Factor Graphs

While Bayes nets provide a modeling framework that allows us to make explicit the *causal* relationships between observed and latent variables, representing the same model as a *factor graph* is more convenient for discussion of inference algorithms.

A factor graph $\mathcal{G} \triangleq \{\mathcal{V}, \mathcal{E}\}$ is a bipartite graph that represents the factorization of a function. A factor graph has vertices $\mathcal{V}$ consisting of factors $f$ and variables $V$.

A factor $f_i$ simply represents a function that takes as an argument its neighbors $\mathbf{V}_i$ (an edge exists between each $f_i$ and all $V_i$ in the set $\mathbf{V}_i$). The entire factor graph can be thought of as representing a single function:

$$f(\mathbf{V}) = f_1(\mathbf{V}_1)f_2(\mathbf{V}_2)\dots f_n(\mathbf{V}_n). \tag{2.17}$$

A factor graph need not represent a normalized probability distribution. Rather, in the context of inference we typically let the factor graph represent the unnormalized joint probability over $\mathbf{V}$:

$$p(\mathbf{V}) \propto \prod_{i=1}^{n} f_i(\mathbf{V}_i). \tag{2.18}$$

Furthermore, in SLAM the variables $\mathbf{V}_i$ represent subsets of $\mathbf{X}$ and $\mathbf{L}$ linked by probabilistic constraints. Formally, we can now write the problem of SLAM with known data association as a factor graph:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto \prod_{i=1}^{n} f_i(\mathbf{V}_i),\ \ \mathbf{V}_i \subseteq \mathbf{X}, \mathbf{L}. \tag{2.19}$$

Figure 2-2 shows an example of a factor graph representing the same distribution as the simple SLAM Bayes net from Figure 2-1

### Tree-Structured Models

In SLAM, the factor graph is constructed from the measurement likelihoods and transition models. Thus, the factorization provided by the model is essentially the same factorization as given in the Bayesian network. The factorization of a joint distribution is not unique, however. For example the following factorizations of distributions

defined for arbitrary elements $a, b, c$ are equally valid:

$$p(a, b, c) = p(a, b \mid c)p(c) \tag{2.20}$$

$$= p(a \mid b, c)p(b, c) \tag{2.21}$$

$$= p(a \mid b, c)p(b \mid c)p(c), \tag{2.22}$$

among others. Thus, while the factor graph gives us a convenient way to write down an inference problem (and can certainly be useful for inference), it may not always be the ideal choice of factorization for inference algorithms.

It is common in many contexts to refactor the graph into a tree structure. Different manifestations of this tree-structured refactorization have been called the *junction tree* in artificial intelligence literature [31] or the *Bayes tree* in recent navigation literature [29]. [11] gives a detailed treatment of the Bayes tree and the *variable elimination algorithm* for performing the refactorization. For brevity, we state only the main results.

The joint distribution $p(\mathbf{V})$ can be refactored into a tree-structured set of cliques $\mathbf{C}_k = \{\mathbf{F}_k, \mathbf{S}_k\}$, where $\mathbf{F}$ and $\mathbf{S}$ are subsets of $\mathbf{V}$ and are termed the "frontal" and "separator" variables, respectively. The resulting factorization can be written as:

$$p(\mathbf{V}) \propto \prod_k p(\mathbf{F}_k \mid \mathbf{S}_k). \tag{2.23}$$

In particular, the frontal and separator variables of a clique $\mathbf{C}_k$ and its parent clique $\mathbf{\Pi}_k$ are related by the following definitions:

$$\mathbf{S}_k \triangleq \mathbf{C}_k \cap \mathbf{\Pi}_k \tag{2.24}$$

$$\mathbf{F}_k \triangleq \mathbf{C}_k \setminus \mathbf{S}_k. \tag{2.25}$$

The root clique $\mathbf{C}_r$, then, is defined as having only frontal variables $\mathbf{F}_r$ and therefore corresponds to a prior on the root variables, $p(\mathbf{F}_r)$.

Throughout this thesis we make use of tree-structured graphical models for in-

ference without much explicit reference to the underlying data structures used. In particular, in Chapter 4 we use the incremental smoothing and mapping framework iSAM2 [30] and in Chapter 5 we leverage the "multimodal" incremental smoothing and mapping work mm-iSAM [20].

### 2.2.3 Algorithms for Variable Elimination

Given graphical models of the form described in the previous section, we now consider the problem of inferring the distribution over latent variables given the observed variables. We consider two algorithms for marginalization: *sum-product* and *max-product*.

The sum-product algorithm, summarized in Algorithm 1 provides a method for exactly recovering the marginal distribution over a subset of variables. Given a set of factors $\mathbf{\Phi}$ and a continuous variable $S$ to eliminate, the sum-product algorithm for variable elimination first computes the product over all factors related to $S$, denoted $\boldsymbol{\phi}$. The result is a function $\psi(\cdots, S)$ of $S$ and all other variables related to $S$ by factors in $\boldsymbol{\phi}$. Finally, the term $\psi(\cdots, S)$ is integrated with respect to $S$, producing the marginal term $\tau(\cdots)$, which is strictly a function of variables other than $S$.

**Algorithm 1** The sum-product elimination algorithm for continuous variables.

$\mathbf{\Phi}$ Set of factors
$S$ Variable to eliminate
$\boldsymbol{\phi} \leftarrow \{\phi \in \mathbf{\Phi}, S \in \text{Scope}(\phi)\}$
$\boldsymbol{\varphi} \leftarrow \mathbf{\Phi} \setminus \boldsymbol{\phi}$
$\psi(\cdots, S) \leftarrow \prod_{\phi \in \boldsymbol{\phi}} \phi(\cdots, S)$
$\tau(\cdots) \leftarrow \int \psi(\cdots, s) ds$
**return** $\tau(\cdots)$

The max-product algorithm, summarized in Algorithm 2 provides a method for exactly recovering the max-marginal for a subset of variables, and is used in the context of maximum *a posteriori* (MAP) inference. As with the sum-product algorithm, we first compute the product over factors adjacent to the variable $S$ to be eliminated. In the max-product algorithm, however, we maximize over possible assignments to $S$, rather than integrating over the domain of $S$, to produce $\tau(\cdots)$. The resulting

marginal is termed a "max-marginal" and is used to recover the maximum *a posteriori* assignments to a set of variables [31].

**Algorithm 2** The max-product elimination algorithm.

$\boldsymbol{\Phi}$ Set of factors
$S$ Variable to eliminate
$\boldsymbol{\phi} \leftarrow \{\phi \in \boldsymbol{\Phi}, S \in \text{Scope}(\phi)\}$
$\boldsymbol{\varphi} \leftarrow \boldsymbol{\Phi} \setminus \boldsymbol{\phi}$
$\psi(\cdots, S) \leftarrow \prod_{\phi \in \boldsymbol{\phi}} \phi(\cdots, S)$
$\tau(\cdots) \leftarrow \max_S \psi(\cdots, S)$
**return** $\tau(\cdots)$

### 2.2.4 MAP Inference as Optimization

Thus far our graphical model formulation of the SLAM problem makes no specific distributional assumptions. That said, in Section 2.2.1 we described a number of nonlinear Gaussian models we are concerned with in this thesis. Indeed, the most commonly used sensor models treat measurements as a nonlinear, but deterministic function of state variables (such as poses and landmarks) corrupted by additive Gaussian noise. It is pertinent to consider the consequences these assumptions have on our graphical model formulation and what new algorithms can be developed leveraging these constraints. In particular, we consider how the problem changes when we aim to recover only the most probable, i.e. the maximum *a posteriori* estimate of vehicle poses and landmarks under such assumptions:

$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}). \tag{2.4, revisited}$$

Consider a general factor graph representation of a joint distribution over poses and landmarks:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) = \prod_i f_i(\mathbf{V}_i), \;\; \mathbf{V}_i \subseteq \{\mathbf{X}, \mathbf{L}\}. \tag{2.26}$$

As before, we denote each factor as $f_i$ and adopt the notation $\mathbf{V}_i$ for the subset of poses and landmarks adjacent to the $i$-th factor. Next, let's adopt the assumption

that each factor $f_i$ is a Gaussian likelihood with respect to some nonlinear function $h_i$ of $\mathbf{V}_i$, that is

$$f_i(\mathbf{V}_i) = \frac{1}{\sqrt{2\pi \det \Sigma_i}} \exp\left\{-\frac{1}{2}\|h_i(\mathbf{V}_i) - \mu_i\|^2_{\Sigma_i}\right\} \tag{2.27}$$

Then, substituting (2.27) into (2.26), we obtain the following expression for the joint belief:

$$\begin{aligned} p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) &= \prod_i f_i(\mathbf{V}_i),\ \mathbf{V}_i \subseteq \{\mathbf{X}, \mathbf{L}\}. && \text{(2.26 revisited)} \\ &= \prod_i \frac{1}{\sqrt{2\pi \det \Sigma_i}} \exp\left\{-\frac{1}{2}\|h_i(\mathbf{V}_i) - \mu_i\|^2_{\Sigma_i}\right\} \\ &\propto \prod_i \exp\left\{-\frac{1}{2}\|h_i(\mathbf{V}_i) - \mu_i\|^2_{\Sigma_i}\right\} && (2.28) \end{aligned}$$

Thus we obtain the result in (2.28) that the posterior belief $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ is, up to a constant of proportionality which depends only on the measurements (and not on our poses $\mathbf{X}$ or landmarks $\mathbf{L}$), equal to the product of exponential terms. Since our concern is to recover the most probable assignment to poses and landmarks, this constant of proportionality has no influence on the result. Following this line of reasoning, we can instead maximize the log of the posterior, which gives:

$$\begin{aligned} \operatorname*{argmax}_{\mathbf{X},\mathbf{L}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) &= \operatorname*{argmax}_{\mathbf{X},\mathbf{L}} \log p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \\ &= \operatorname*{argmax}_{\mathbf{X},\mathbf{L}} \log \prod_i \exp\left\{-\frac{1}{2}\|h_i(\mathbf{V}_i) - \mu_i\|^2_{\Sigma_i}\right\} \\ &= \operatorname*{argmax}_{\mathbf{X},\mathbf{L}} \sum_i -\frac{1}{2}\|h_i(\mathbf{V}_i) - \mu_i\|^2_{\Sigma_i} \\ &= \operatorname*{argmin}_{\mathbf{X},\mathbf{L}} \sum_i \|h_i(\mathbf{V}_i) - \mu_i\|^2_{\Sigma_i}. && (2.29) \end{aligned}$$

The result in (2.29) has the form of a nonlinear least-squares problem in terms of the Mahalanobis distance. In order to optimize the resulting expression in (2.29) we can use nonlinear optimization methods like Gauss-Newton and Levenberg-Marquadt.

Such methods depend on the *linearization* of the $h_i$'s.

Specifically, we consider a first-order Taylor approximation of each $h_i$ given a particular linearization-point consisting of an assignment to the elements of $\mathbf{V}_i$. Let the linearization point be $\hat{\mathbf{V}}_i$, then the Taylor expansion of of $h_i$ can be written as:

$$h_i(\mathbf{V}_i) \approx h_i(\hat{\mathbf{V}}_i) + \nabla h_i(\hat{\mathbf{V}}_i)(\mathbf{V}_i - \hat{\mathbf{V}}_i), \tag{2.30}$$

where $\nabla g(\hat{\boldsymbol{v}})$ denotes the Jacobian of a function $g : \mathbb{R}^n \to \mathbb{R}^m$ evaluated at $\hat{\boldsymbol{v}}$ with element $(i, j)$ defined as:

$$\nabla g(\hat{\boldsymbol{v}})_{(i,j)} \triangleq \left.\frac{\partial g_i}{\partial \boldsymbol{v}_j}\right|_{\hat{\boldsymbol{v}}_j}. \tag{2.31}$$

One caveat of the above is that we have assumed vector-valued $\mathbf{V}_i$ and $h_i(\mathbf{V}_i)$, where in general this is not the case; poses $\boldsymbol{x}$ are elements of the two- or three-dimensional Euclidean groups, and similarly measurements may have their own manifold structure. We will not discuss the subject of optimization on manifolds in detail in this thesis, but refer to [11] for a more complete treatment of manifold optimization for SLAM.

Substituting the Taylor expansion of $h_i$ from (2.30) into the nonlinear least-squares optimization from (2.29), we obtain:

$$\|h_i(\mathbf{V}_i) - \mu_i\|^2_{\Sigma_i} \approx \|h_i(\hat{\mathbf{V}}_i) + \nabla h_i(\hat{\mathbf{V}}_i)(\mathbf{V}_i - \hat{\mathbf{V}}_i) - \mu_i\|^2_{\Sigma_i}, \tag{2.32}$$

which allows us to rewrite the original problem as a weighted linear least-squares problem. We make this explicit as follows by defining the matrix $J_i$ as the Jacobian of the $i$-th measurement model $h_i$ at the linearization point $\hat{\mathbf{V}}_i$:

$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmin}} \sum_i \|h_i(\hat{\mathbf{V}}_i) + J_i(\mathbf{V}_i - \hat{\mathbf{V}}_i) - \mu_i\|^2_{\Sigma_i} \tag{2.33}$$

Defining the error residual $r_i(\mathbf{V}_i) = h_i(\hat{\mathbf{V}}_i) + J_i(\mathbf{V}_i - \hat{\mathbf{V}}_i) - \mu_i$, we can rewrite the expression (2.33) in terms of the L2-norm (and therefore as a standard linear least-

squares problem) as follows:

$$\begin{aligned}\|h_i(\hat{\mathbf{V}}_i) + J_i(\mathbf{V}_i - \hat{\mathbf{V}}_i) - \mu_i\|^2_{\Sigma_i} &= \|r_i(\mathbf{V}_i)\|^2_{\Sigma_i} \\ &= \|\Sigma^{-1/2} r_i(\mathbf{V}_i)\|^2_2. \end{aligned} \tag{2.34}$$

Thus, by changing variables to $A_i = \Sigma_i^{-1/2} J_i$ and $b_i = \Sigma_i^{-1/2}(\mu_i - h_i(\hat{\mathbf{V}}_i))$, we obtain a simple linear least-squares problem:

$$\begin{aligned}\hat{\mathbf{X}}, \hat{\mathbf{L}} &= \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmin}} \sum_i \|A_i(\mathbf{V}_i - \hat{\mathbf{V}}_i) - b_i\|^2_2 \\ &= \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmin}} \|\mathbf{A}(\mathbf{V} - \hat{\mathbf{V}}) - \boldsymbol{b}\|^2_2, \end{aligned} \tag{2.35}$$

where the last line (2.35) follows from block-wise stacking each term in the sum.

Broadly, nonlinear optimization methods such as Gauss-Netwon and Levenberg-Marquadt perform linearization as we have shown above. They then solve for the term $\mathbf{V} - \hat{\mathbf{V}}$ in the linearized problem from which values of $\mathbf{X}$ and $\mathbf{L}$ can be recovered. This process is repeated until the solution converges, or some fixed maximum number of iterations is reached.

Lastly, the linearized joint covariance matrix $\boldsymbol{\Sigma}$ can be obtained, using the optimal values of poses and landmarks $\hat{\mathbf{X}}, \hat{\mathbf{L}}$ as the linearization point, as:

$$\boldsymbol{\Sigma} = (\mathbf{A}^T \mathbf{A})^{-1}. \tag{2.36}$$

For this reason, $\mathbf{A}$ is referred to as the "square-root information matrix"; its square corresponds to the information matrix, i.e. the inverse of the covariance matrix. The above equality is straightforward to show by considering the definition of $\mathbf{A}$ as $\boldsymbol{\Sigma}^{-1/2}\mathbf{J}$ where $\mathbf{J}$ is simply the concatenation of each measurement Jacobian $J_i$. In practice this multiplication is not performed directly to obtain the covariance. Rather, more efficient methods have been presented, such as those in [28].

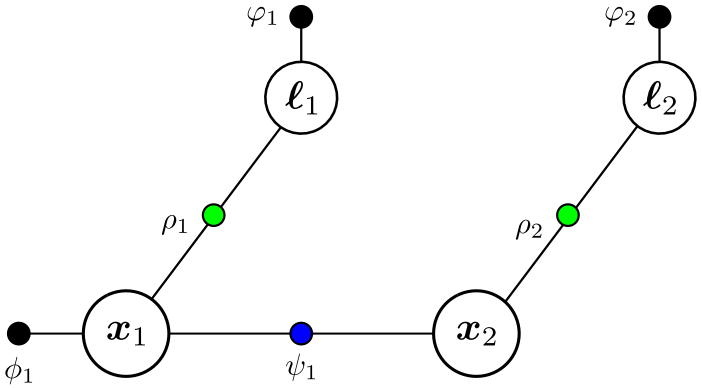



Figure 2-2: When we can uniquely identify the landmark corresponding to each measurement, the only variables to infer are robot poses and landmarks. This factor graph represents such a case, where we have two vehicle poses and two landmarks. We have a prior factor $\phi_1$ on the initial vehicle pose, and priors $\varphi_1, \varphi_2$ on each of the landmark locations. An odometry measurement $\psi_1$ links the poses, and range-bearing measurements $\rho_1$ and $\rho_2$ are made from the first and second poses to the first and second landmarks, respectively.

## 2.3 Summary

In this section we described several preliminaries related to state-of-the-art SLAM systems used in this work. In particular, we describe the SLAM problem as one of Bayesian inference, defined by probabilistic measurement models. We discussed the measurement models relevant to the work proposed in this thesis, including odometry models, range and bearing models, and semantic measurement models. We described the sum-product and max-product algorithms for inference in graphical models without assumptions on the distributional form of measurement models. Lastly, we showed how the MAP inference problem for Gaussian factor graphs can be re-cast as one of nonlinear optimization. The algorithms discussed form the foundation for state-of-the-art SLAM methods, with the former being relevant to non-Gaussian SLAM in the contexts of MAP inference and full posterior inference, and the latter being the backbone for modern SLAM systems that assume Gaussian noise.

# Chapter 3

# Data Association

## 3.1 Overview

In the previous section, discussion was limited to the case in which data associations—the correspondences between sensor measurements and environmental landmarks—are assumed to be known. In the probabilistic framework we developed in the previous chapter, the likelihood of a measurement (and consequently the belief over poses and landmarks given our measurements) depends on which landmark caused the measurement. Similarly, in this section we describe how the determination of the data association itself critically depends on our belief over robot poses and landmarks. In this sense, as with the general SLAM problem, this presents a chicken and egg problem: we need measurement-landmark correspondences to infer poses and landmarks (i.e. to do SLAM), but we need an estimate of poses and landmarks to infer data associations.

In this chapter, we first outline the above problem of joint inference of poses, landmarks, and data associations. We then describe a variety of methods for addressing the data association problem. Broadly, data association methods may be *lazy* and defer making association decisions (and allow for revision of previous decisions), or *proactive*, making a decision about data associations (or their respective probabilities) at a given point in time using only the information prior to that moment, and doing so without revision. While this thesis concerns proactive approaches, many of the

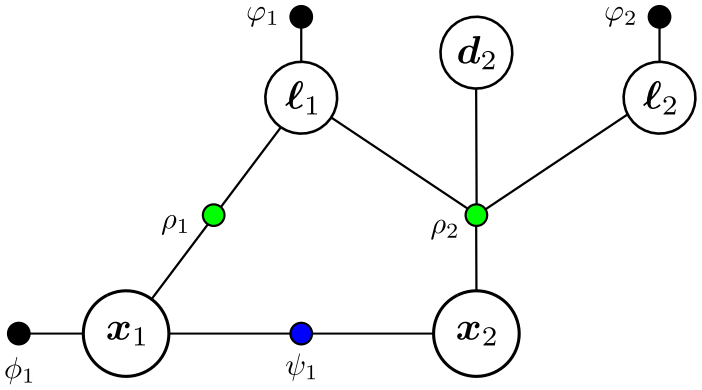



Figure 3-1: When data associations are unknown, they introduce an additional latent variable into the factor graph. This graphical model extends the SLAM problem from Figure 2-2 to a situation where the data association for the second range-bearing measurement $\rho_2$ is unknown.

ideas we develop are applicable in the general setting of lazy data association as well. We conclude by discussing related works in semantic SLAM and data association.

## 3.2 SLAM with Unknown Data Association

When correspondences between measurements and landmarks are not known *a priori*, they must also be inferred. Let $d_t^k$ denote a data association for measurement $k$ taken at pose $\boldsymbol{x}_t$, such that $d_t^k = j$ signifies that measurement $\boldsymbol{z}_t^k$ corresponds to landmark $\boldsymbol{\ell}_j$. Let $\mathbf{D} \triangleq \{\boldsymbol{d}_t\}_{t=1}^T$ denote the set of all associations of measurements to landmarks. The data associations at each time $\boldsymbol{d}_t$ now appear as additional latent variables in the factor graph, as in Figure 3-1.

We now obtain a new factor graph characterization of the joint:

$$p(\mathbf{X}, \mathbf{L}, \mathbf{D} \mid \mathbf{Z}) = \prod_i f_i(\mathbf{V}_i), \;\; \mathbf{V}_i \subseteq \{\mathbf{X}, \mathbf{L}, \mathbf{D}\}. \tag{3.1}$$

More concretely, in the case of the example in Figure 3-1, the joint posterior can be

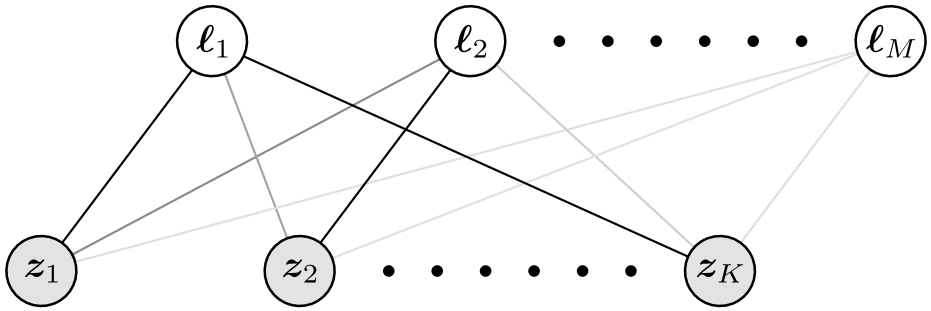



Figure 3-2: At a given time $t$, there are $K$ measurements that must be associated to any of $M$ landmarks. Depicted above is a graphical representation of this problem. Here we denote the probability that a measurement arose due to the observation of a particular landmark as the darkness of the edge that connects the nodes. Whether the association process is done for each measurement independently or with consideration of all $K$ measurements simultaneously is the distinction between individual compatibility (IC) and joint-compatibility (JC) approaches [28].

written:

$$\begin{aligned} p(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{d}_2 \mid \mathbf{Z}) = \phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2) \times \dots \text{ (Prior terms)} \\ \dots \psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2) \times \dots \text{ (Odometry meas.)} \\ \dots \rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1)\rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{d}_2) \text{ (Landmark meas.).} \end{aligned}$$

Notably, factors are now functions not only of poses and landmarks, but also association variables. The factor $\rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{d}_2)$ corresponds to the measurement likelihood conditioned on the association $\boldsymbol{d} = j$ with $j \in \{1, 2\}$. More concretely, it corresponds to the measurement model $p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{d}_2 = j) = p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\ell}_j)$.

### 3.2.1 Measurement Gating

A problem that arises when data associations are unknown is when a measurement corresponds to a landmark outside the known set. Often it is the case in SLAM that we do not know how many landmarks exist in the environment *a priori*, and we must determine online whether a measurement corresponds to a new landmark $\boldsymbol{\ell}_{new}$. The most common and simple method for determining whether a measured landmark is novel is *gating*. Gating refers to placing a threshold $\tau$ on the probability

that a measurement arose from each of the candidate landmarks. If the probability of an association to every currently known landmark falls below the threshold $\tau$, we assign a new landmark to the measurement. In the case of linearized Gaussian models, computing data association probabilities is equivalent to placing a Mahalanobis distance threshold [28]. We discuss specific methods for computing data association probabilities under Gaussian distribution assumptions and for arbitrary distributions in Chapters 4 and 5 respectively.

### 3.2.2 Problem Dimensionality

Perhaps the most critical problem encountered in the consideration of unknown data association is the dimensionality of the problem. When data associations are known, the dimensionality of the SLAM joint is related directly to the product of the number of poses $T$ and the number of landmarks $M$, which is in the worst case equal to the number of measurements $|\mathbf{Z}|$. To simplify the analysis, recall that $K_t$ is the number of measurements in $\boldsymbol{z}_t$ (the set of measurements taken at time $t$) for all $t$, and define $K = \max_t K_t$. The cardinality of $\mathbf{Z}$, then, is in $\mathcal{O}(KT)$. Since generally we have $K << T$, i.e. the maximum number of measurements at a particular time $t$ is much less than the total number of discrete time-steps in the entire history of a robot trajectory, we can think of this set as growing linearly in $T$.

When we consider unknown data associations, the space of the joint distribution is multiplied by $|\mathbb{D}|$, the size of the space of all possible measurement-landmark correspondences. For each measurement $\boldsymbol{z}$, there are at most $|\mathbf{Z}|$ candidate associations. By our previous analysis, we have $|\mathbf{Z}| \in \mathcal{O}(T)$, Thus, a worst-case analysis suggests that $|\mathbb{D}| \in \mathcal{O}(T^T)$, i.e. exponential in the number of discrete time-steps. As a consequence, it is generally not feasible to enumerate every possible set of data associations. Fortunately, a number of reasonable approaches that mitigate the exponential growth in data associations are available. We can leverage, for example, the property that though the space of data associations is immense, very few of the outcomes $\mathbf{D} \in \mathbb{D}$ have non-negligible probability. This property allows us to circumvent joint consideration of all possible data associations at once. As an example, for detections of

landmarks in a given image, we need not consider associations to landmarks which are with very high probability out of the image frame. We make note of approximations like these when we make use of them in Chapters 4 and 5, but the subsequent discussion of data association methods neglects the computational burden of enumerating the space of possible correspondences. Rather, we assume approximations can be made that make this enumeration tractable and focus on the computational issues associated with inference.

## 3.3 Maximum-Likelihood Data Association

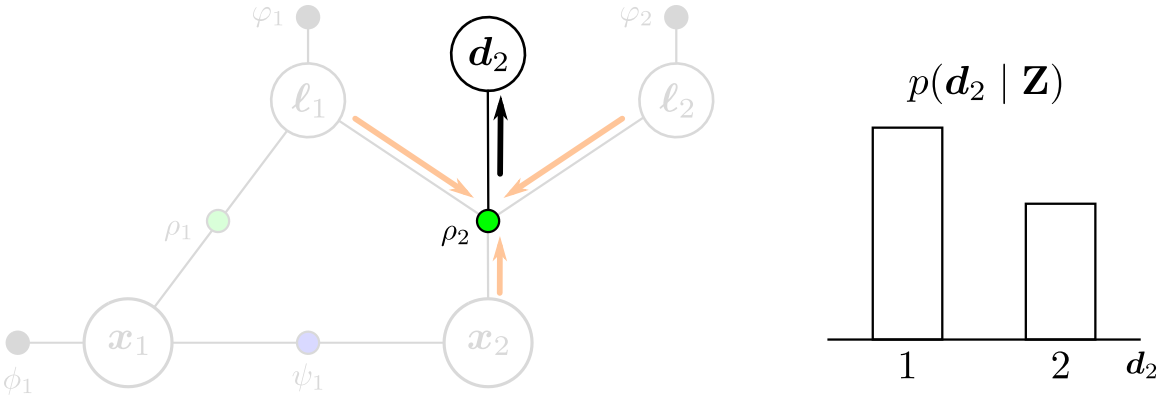
One of the most common solutions to the data association problem is maximum-likelihood estimation. That is, given an initial estimate of poses and landmarks $\mathbf{X}^{(0)}$ and $\mathbf{L}^{(0)}$, respectively, maximum-likelihood data association performs the following optimization:

$$\hat{\mathbf{D}} = \operatorname*{argmax}_{\mathbf{D}} p(\mathbf{D} \mid \mathbf{X}^{(0)}, \mathbf{L}^{(0)}, \mathbf{Z}) \tag{3.2}$$

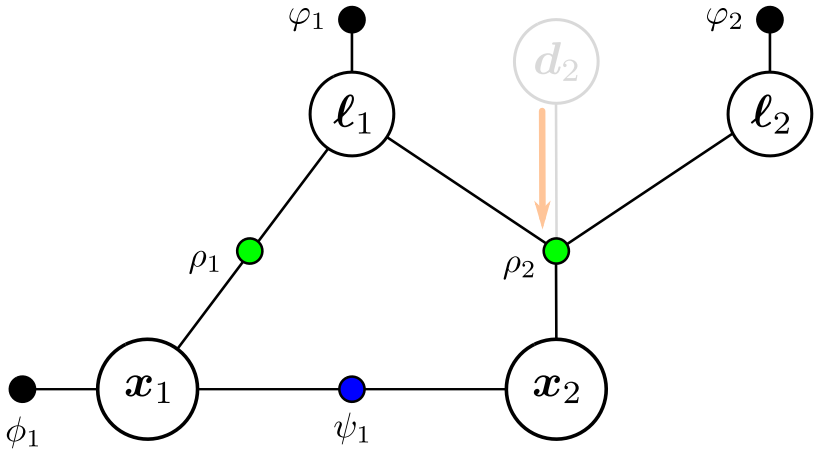
$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \operatorname*{argmax}_{\mathbf{X},\mathbf{L}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}, \hat{\mathbf{D}}). \tag{3.3}$$

Often the pose and landmark estimates $\mathbf{X}^{(0)}$ and $\mathbf{L}^{(0)}$ are produced using a subset of measurements and marginalized out to compute the data association probabilities. For example, in the case of proactive data association, usually the pose and landmark estimates are obtained using measurements $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t-1}$, and the pose and landmark variables are marginalized out (a process we describe in detail for the Gaussian and non-Gaussian cases in Chapters 4 and 5, respectively).

Broadly, in the maximum-likelihood approach, data associations are computed and fixed, then the SLAM solution is optimized assuming the fixed set of data associations. Methods like the Hungarian algorithm [32] and joint compatibility branch and bound [40] are typically used to simultaneously compute an optimal assignment of all measurements $\boldsymbol{z}_t^k,\ k = 1, \ldots, K_t$ observed at a pose $\boldsymbol{x}_t$ to landmarks. While very efficient and easy to implement, this method can be brittle. In particular, if there is




(a) Computing belief over a data association variable by marginalizing out the history of poses and landmarks (left). This produces the marginal belief over data association candidates (right).




(b) Marginalization of the data association variable produces an equivalent factor graph with a non-Gaussian factor linking $\boldsymbol{x}_2$, $\boldsymbol{\ell}_1$, and $\boldsymbol{\ell}_2$.

Figure 3-3: Computing the belief over possible data associations by marginalizing out the current poses and landmarks.

an error in a data association, there is no opportunity to revise the association. In approaches that address SLAM using nonlinear least-squares optimization, there is a quadratic cost associated with each term. An incorrect association can give otherwise correct solutions a high cost, which often makes the minimum cost solution very far from the ground-truth solution (see for example the case of maximum-likelihood associations in Figure 1-2).

## 3.4 Probabilistic Data Association

An alternative solution is to consider *probabilistic* data associations. If we had access to the probability of each data association, we could marginalize out data associations when computing the solution to the SLAM problem,

$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}} \sum_{\mathbf{D}} p(\mathbf{X}, \mathbf{L}, \mathbf{D} \mid \mathbf{Z}) \tag{3.4}$$

$$= \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}} \sum_{\mathbf{D}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) p(\mathbf{D} \mid \mathbf{Z}) \tag{3.5}$$

$$= \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}} \mathbb{E}_{\mathbf{D}} \left[ p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) \mid \mathbf{Z} \right] \tag{3.6}$$

The approximate marginal distribution represented by the expectation over data association hypotheses—even when the individual measurement likelihoods are well-represented by Gaussian distributions—is almost always multimodal in practice. This may not be immediately obvious, however. Note that each term $p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z})$ represents the posterior belief over poses and landmarks given a fixed set of data associations. When measurement models are Gaussian conditioned on data associations, this posterior is Gaussian, as we have seen in the previous chapter. Each term $p(\mathbf{D} \mid \mathbf{Z})$ is simply a scalar probability. Thus the resulting expectation takes the form of a sum-mixture of Gaussians.

The sum-mixture of Gaussians form is outside the realm of traditional least-squares approaches to SLAM. Beyond this issue, as we have previously mentioned, the space $\mathbb{D}$ is in the worst case exponentially large in the number of measurements. A number of solutions exist that maintain a set of Gaussian solutions that branch with each set of new hypothesis. Work in the this area has primarily focused on methods to prune the space of plausible hypotheses (e.g. [9], [50], and recently [56]), leveraging the fact that $p(\mathbf{D} \mid \mathbf{Z})$ is sparse over $\mathbb{D}$.

A recent solution making use of expectation-maximization iterates between com-

puting the data association probabilities and the conditional log-likelihood [5]:

$$\hat{\mathbf{X}}^{(i+1)}, \hat{\mathbf{L}}^{(i+1)} = \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}} \mathbb{E}_{\mathbf{D}} \left[ \log p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}, \mathbf{D}) \mid \mathbf{X}^{(i)}, \mathbf{L}^{(i)}, \mathbf{Z} \right]. \tag{3.7}$$

The effect of this approach is that the sum of Gaussians in the sum-marginal is replaced by a geometric mean. This has the benefit of preserving the Gaussian posterior assumption, and iterating in this fashion provides guaranteed convergence. Throughout this thesis we refer to this method as *Gaussian probabilistic data association (Gaussian PDA)*, as it performs probabilistic data association while preserving the Gaussian distribution assumptions. This approach solves for point estimates of poses and landmarks at each iteration using an approach similar to weighted maximum likelihood estimation [57], where weights are determined by estimated data association probabilities. In practice, recomputing data association probabilities for all previous measurements is a computational burden, so typically data association probabilities are computed once proactively (i.e. after each keyframe), resulting in solutions somewhere *between* the modes induced by the plausible association hypotheses. Nonetheless, approximate methods for computing the permanent of a matrix exist which have been shown to permit more efficient recomputation of data association probabilities [2].

### 3.4.1 Our Approach

We consider two approaches in this thesis for addressing non-Gaussianity in the semantic SLAM problem with probabilistic data associations. Both leverage the idea of marginalizing out the discrete data association variables in the solution to the SLAM problem, as described graphically in Figure 3-3. The sum-marginalization we initially proposed in Equation (3.4) computes the posterior belief $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ before maximizing. In our first approach, described in Chapter 4, we show that if our goal is solely MAP inference, we can replace the sum-marginal with a max-marginal and perform (in principle) exact MAP inference in a nonlinear least-squares framework.

In many cases, however, the MAP estimate is insufficient. For example, if we

aim to inform planning decisions based on uncertainty in the posterior (as is the case broadly in *active SLAM*), richer posterior representations are required. In Chapter 5, we consider directly inferring the (non-Gaussian) posterior $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ represented by the marginalization in Equation (3.4). The resulting computation is necessarily approximate, but provides a rich representation for planning that incorporates ambiguity due to uncertain data associations. Furthermore, this approach is capable of coping with non-Gaussian uncertainties in measurement likelihoods, such as those arising from nonlinearity in measurements, physical and environmental effects like acoustic multipath, and undetermined systems (as encountered in range-only and bearing-only inference problems), topics we will revisit in Chapter 6.

# Chapter 4

# Robust Semantic SLAM with Max-Mixtures

In the previous chapter, we outlined several approaches to inference in graphical models and introduced the two core algorithms from which the methods in this thesis can be derived. In this chapter, we are specifically concerned with maximum *a posteriori* MAP inference for semantic SLAM with unknown data association. That is, we want to find the most probable set of robot poses and landmarks given the measurements:

$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}). \qquad \text{(2.4, revisited)}$$

In this chapter, we present an algorithm addressing the MAP inference problem for ambiguities that can be represented by a mixture of Gaussians. In particular, we make use of max-product marginalization, as described in Algorithm 2 to eliminate discrete data association variables in the SLAM solution. In doing so, we arrive at a formulation that is similar to the "max-mixtures" method of Olson and Agarwal [42].

## 4.1 Max-Marginalization of Data Associations

As described in the previous chapter, we approach the problem of SLAM with unknown data association by introducing the associations as additional latent variables.

In order to address the specific problem of MAP inference posed in (2.4), we would like to eliminate the influence of the data association variable. Previously we introduced two algorithms for variable elimination: the sum-product algorithm and max-product algorithm. For MAP inference it will suffice to compute the so-called "max-marginal" of $\mathbf{X}$ and $\mathbf{L}$ over all possible $\mathbf{D}$. That is:

$$\hat{\mathbf{X}}, \hat{\mathbf{L}} = \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \tag{2.4, revisited}$$

$$= \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}} \left[ \max_{\mathbf{D}} p(\mathbf{X}, \mathbf{L}, \mathbf{D} \mid \mathbf{Z}) \right] \tag{4.1}$$

$$= \underset{\mathbf{X},\mathbf{L}}{\operatorname{argmax}} \left[ \max_{\mathbf{D}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) p(\mathbf{D} \mid \mathbf{Z}) \right]. \tag{4.2}$$

We refer to the inner maximization term as the "max-marginal" of $\mathbf{X}$ and $\mathbf{L}$, formally defined as:

$$\mu(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) = \max p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) p(\mathbf{D} \mid \mathbf{Z}), \tag{4.3}$$

where we use the notation $\mu(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ to distinguish this function from the posterior belief $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z})$ (which would be equal to the result of sum-marginalization of data associations). Under common Gaussian measurement assumptions. This result conveniently replaces the sum-of-Gaussians representation of the true posterior with a maximization over terms which individually are Gaussian.

### 4.1.1 Proactive Max-Marginalization

Exact computation of the max-marginal over all possible data associations in Equation (4.2) is computationally expensive. In the Gaussian case, and when applying nonlinear least-squares optimization algorithms like Gauss-Newton and Levenberg-Marquadt, we need to evaluate this max-marginal each time we update the assignment of $\mathbf{X}$ and $\mathbf{L}$ (i.e. on every iteration of the optimization procedure). As we described in Chapter 3, the exponentially large set of possible data associations $\mathbb{D}$ makes brute-force consideration of ever association impractical. This becomes par-

ticularly relevant when we consider that an autonomous robot requires up-to-date estimates of its own state and the state of the environment in order to reliably navigate. As a consequence, we perform data association proactively. This allows us to eliminate data associations as new measurements are made, and the set of plausible associations to a single measurement (or subset of measurements) at a particular time is ostensibly smaller than the set of all possible associations to measurements at any time. However, this method is approximate in that it does not consider the influence of future measurements when computing data association probabilities.

In particular, suppose we have some set of previous measurements $\mathbf{Z}^-$ and new measurements $\mathbf{Z}^+$ (i.e. $\mathbf{Z} = \mathbf{Z}^+ \cup \mathbf{Z}^-$ with $\mathbf{Z}^+ \cap \mathbf{Z}^- = \emptyset$). We aim to compute the max-marginal over associations to the new measurements, denoted $\mathbf{D}^+$. Formally, we have the following:

$$\begin{aligned} p(\mathbf{X}, \mathbf{L}, \mathbf{D}^+ \mid \mathbf{Z}^+, \mathbf{Z}^-) &= \frac{p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+, \mathbf{Z}^-) p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}^+, \mathbf{Z}^-) p(\mathbf{D}^+ \mid \mathbf{Z}^-) p(\mathbf{Z}^-)}{p(\mathbf{Z}^+, \mathbf{Z}^-)} \\ &\propto p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+) p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-) p(\mathbf{D}^+ \mid \mathbf{Z}^-), \end{aligned} \tag{4.4}$$

where in the last line we have removed the proportionality constant $p(\mathbf{Z}^-)/p(\mathbf{Z}^+, \mathbf{Z}^-)$, used the conditional independence $p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+, \mathbf{Z}^-) = p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+)$, and used the conditional independence $p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}^+, \mathbf{Z}^-) = p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$, since $\mathbf{D}^+$ consists of associations to only measurements outside of $\mathbf{Z}^-$. Using the above decomposition, we apply max-marginalization to data associations:

$$\mu(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^+, \mathbf{Z}^-) = \max_{\mathbf{D}^+} \left[ p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+) p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-) p(\mathbf{D}^+ \mid \mathbf{Z}^-) \right] \tag{4.5}$$

$$= p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-) \max_{\mathbf{D}^+} \left[ p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+) p(\mathbf{D}^+ \mid \mathbf{Z}^-) \right]. \tag{4.6}$$

Here $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$ is the (potentially non-Gaussian) posterior distribution over poses and landmarks after sum-marginalization of data associations to the measurements $\mathbf{Z}^-$. For the purposes of optimization in the Gaussian case, we take this as the max-marginal $\mu(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$.

As an example, consider a simple factor graph consisting of two vehicles poses

with an odometry constraint between them, each measuring one of two candidate landmarks, each with a prior (as in Figures 3-1 and 3-3). We can write the posterior as:

$$p(\mathbf{X}, \mathbf{L}, \mathbf{D} \mid \mathbf{Z}) \propto \psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2)\rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_1)\rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_2)\phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2),$$

where here $\mathbf{D} = \{d_1, d_2\}$. Suppose that each distribution is Gaussian conditioned on the data association variables $d_1$ and $d_2$; thus the posterior conditioned on the associations is jointly Gaussian (subject to the appropriate linearization of any non-linear models). Straightforward sum-product marginalization of the data association variables (according to Algorithm 1) here will result in a decidedly non-Gaussian inference problem. Furthermore, while the non-Gaussian problem represented here is tractable (as there are only 4 possible solutions), exact inference for these types of problems is generally computationally intractable due to the large number of modes that arise in the posterior. Max-marginalization (which retains the Gaussianity of the problem conditioned on the data associations) can be computed as:

$$\mu(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto \max_{\mathbf{D} \in \mathbb{D}} \psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2)\rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_1)\rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_2)\phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2). \quad (4.7)$$

Assuming that data associations are independent (conditioned on previous measurements) we may in fact push the maximization into the individual terms. In general, the following reduction applies to any two factors $\phi_1$ and $\phi_2$:

$$\max_x(\phi_1 \cdot \phi_2) \equiv \phi_1 \cdot \max_x \phi_2. \quad (4.8)$$

whenever $x \notin \text{Scope}(\phi_1)$ [31]. Thus, when data associations can be considered conditionally independent given previous measurements (as we have written them in the

original max-marginal equation), the resulting max-marginal can be rewritten as:

$$\mu(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto$$
$$\psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2)\phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2) \max_{d_1 \in \mathcal{D}_1} \rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_1) \max_{d_2 \in \mathcal{D}_2} \rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_2). \quad (4.9)$$

The consequences of this simple change are quite significant: the complexity of evaluating the max operators in the above expression is in the worst case *quadratic*[1], rather than exponential in the number of measurements. Practically, there will often be many fewer landmarks (and thus candidate hypotheses) than measurements, so in practice this is usually roughly linear in the number of measurements. In practice, of course, computing the data association probability for a measurement requires marginalization of the current belief over poses and landmarks, and that belief hinges critically on previous associations. The result is that we have arrived at an approximate max-product algorithm for SLAM with unknown data associations.

## 4.2 Max-Mixtures Semantic SLAM

To perform covariance recovery for computing data association probabilities, we use the method of Kaess and Daellart [28]. In particular, consider a single measurement of a landmark $\boldsymbol{z}$, which in our case consists of the estimated range and bearing to the landmark (though *which* landmark is unknown *a priori*). Furthermore, let $\Sigma$ denote the block joint covariance matrix:

$$\Sigma = \begin{bmatrix} \Sigma_{tt} & \Sigma_{tj} \\ \Sigma_{jt} & \Sigma_{jj} \end{bmatrix} \quad (4.10)$$

between a pose $\boldsymbol{x}_t$ and candidate landmark $\boldsymbol{\ell}_j$, obtained from the solution to the factor graph obtained at time $t$. We assume the data association probability is pro-

[1]The true total number of operations is roughly the product of the number of hypotheses per measurement and the number of measurements. Here we say this is in the worst-case quadratic, since we could plausibly have at most one landmark per measurement, discounting the possibility of a false positive measurement that arises without the existence of a landmark.

portional to the likelihood $p(\boldsymbol{z}_t \mid \boldsymbol{d}_t, \mathbf{Z}^-)$. We can write the joint distribution between a new measurement and data association conditioned on all previous measurements as follows:

$$p(\boldsymbol{z}_t^k \mid d_t^k = j, \mathbf{Z}^-) = \iint p(\boldsymbol{z}_t^k, \boldsymbol{x}_t, \boldsymbol{\ell}_j \mid d_t^k = j, \mathbf{Z}^-) d\boldsymbol{x}_t d\boldsymbol{\ell}_j \tag{4.11}$$

$$= \iint p(\boldsymbol{z}_t^k \mid d_t^k = j, \boldsymbol{x}_t, \boldsymbol{\ell}_j) p(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-) d\boldsymbol{x}_t d\boldsymbol{\ell}_j. \tag{4.12}$$

With data associations marginalized out, the belief $p(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-)$ would be generally non-Gaussian. Consequently, we again make an approximation. We use the max-marginal for the posterior $\mu(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-)$ evaluated at the current MAP estimate of $\boldsymbol{x}_t$ and $\boldsymbol{\ell}_j$, which we denote $\hat{\mu}(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-)$, to form the tractable approximation:

$$p(\boldsymbol{z}_t^k \mid d_t^k = j, \mathbf{Z}^-) \approx \iint p(\boldsymbol{z}_t^k \mid d_t^k = j, \boldsymbol{x}_t, \boldsymbol{\ell}_j) \hat{\mu}(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-) d\boldsymbol{x}_t d\boldsymbol{\ell}_j. \tag{4.13}$$

Recalling the factored measurement model assumptions from Section 2.2.1, we also have that the distribution of the form $p(\boldsymbol{z} \mid \boldsymbol{d}, \mathbf{Z}^-)$ can be broken into the product $p(\boldsymbol{z}^r \mid \boldsymbol{d}, \mathbf{Z}^-) p(\boldsymbol{z}^b \mid \boldsymbol{d}, \mathbf{Z}^-) p(\boldsymbol{z}^s \mid \boldsymbol{d}, \mathbf{Z}^-)$, where $\boldsymbol{z}^r$, $\boldsymbol{z}^b$, and $\boldsymbol{z}^s$ are the range, bearing, and semantic class measurements, respectively[2]. Consequently, subject to the max-marginal approximation above, we obtain the following equivalent expression:

$$\begin{aligned} p(\boldsymbol{z}_t^k \mid d_t^k = j, \mathbf{Z}^-) &= p(\boldsymbol{z}_t^s \mid d_t^k = j, \mathbf{Z}^-) p(\boldsymbol{z}_t^r, \boldsymbol{z}_t^b \mid d_t^k = j, \mathbf{Z}^-) \\ &\approx \left[ \sum_c p(\boldsymbol{z}_t^s \mid \boldsymbol{\ell}_j^s = c) p(\boldsymbol{\ell}_j^s = c \mid \mathbf{Z}^-) \right] \times \dots \\ &\quad \dots \left[ \iint p(\boldsymbol{z}_t^r, \boldsymbol{z}_t^b \mid d_t^k = j, \boldsymbol{x}_t, \boldsymbol{\ell}_j) \hat{\mu}(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-) d\boldsymbol{x}_t d\boldsymbol{\ell}_j \right]. \end{aligned}$$

We have dropped the superscript $k$ for a single measurement to emphasize the decomposition into range, bearing, and semantic components, but we are considering a single measurement $\boldsymbol{z}_t^k$ throughout. Since the range and bearing measurements are

[2]It is simple to consider range-bearing measurements which are jointly Gaussian, but typical noise model assumptions for such sensors assume diagonal covariances, and therefore assume the same factorization as presented here

geometric and are agnostic to the class of the landmark, the original integral can be broken into a discrete summation over landmark classes and integration over robot pose and landmark location. The result of the summation is simply a scalar. Furthermore, since all of the terms in the integral are Gaussian, it can be simplified as follows:

$$\begin{aligned} p(\boldsymbol{z}_t^r, \boldsymbol{z}_t^b \mid d_t^k = j, \mathbf{Z}^-) &\approx \iint \frac{1}{\sqrt{|2\pi\Gamma|}} e^{-\frac{1}{2}\|h_k(\boldsymbol{x}_t, \boldsymbol{\ell}_j) - \mathbf{z}\|^2_\Gamma} \frac{1}{\sqrt{|2\pi\Sigma|}} e^{\|\mathbf{x} - \hat{\mathbf{x}}\|^2_\Sigma} d\boldsymbol{x}_t d\boldsymbol{\ell}_j \\ &\approx \frac{1}{\sqrt{|2\pi C_{tjk}|}} e^{-\frac{1}{2}\|h_k(\hat{\mathbf{x}}) - \mathbf{z}\|^2_{C_{tjk}}} \end{aligned} \tag{4.14}$$

where $\mathbf{x}$ is the stacked vector representation $[\boldsymbol{x}_t, \boldsymbol{\ell}_j]^T$, $\mathbf{z}$ is the stacked vector $[\boldsymbol{z}_t^r, \boldsymbol{z}_t^b]$, $\hat{\mathbf{x}}$ is the mean of the joint distribution over $\boldsymbol{x}_t$ and $\boldsymbol{\ell}_j$. Furthermore, we take $\Gamma$ to be the joint covariance over $\boldsymbol{z}^r$ and $\boldsymbol{z}^b$. The covariance $C_{tjk}$, then, is defined as:

$$C_{tjk} \triangleq \left.\frac{\partial h_k}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}} \Sigma \left.\frac{\partial h_k}{\partial \mathbf{x}}\right|^T_{\hat{\mathbf{x}}} + \Gamma. \tag{4.15}$$

Taking the distribution $p(\boldsymbol{d} \mid \mathbf{Z}^-)$ as proportional to the above (over all assignments to $\boldsymbol{d}$) and normalizing the result appropriately, we compute the max-mixture factor for a measurement $\boldsymbol{z}_t$ as:

$$f(\boldsymbol{x}_t, \boldsymbol{\ell}_{1:M}) = \max_j p(\boldsymbol{z}_t \mid \boldsymbol{x}_t, \boldsymbol{\ell}_j) p(d_t^k = j \mid \mathbf{Z}^-). \tag{4.16}$$

Here we have used the most general case where we consider all landmarks $j = 1, \ldots, M$, but in practice we use measurement gating as described in Section 3.2.1 to reduce the subset of landmarks that are examined in the maximization.

Finally, we can recover maximum *a posteriori* landmark semantic class estimates (assuming uniform priors) as in [5] follows:

$$\hat{\boldsymbol{\ell}}_j^s = \operatorname*{argmax}_c \prod_t \sum_{\boldsymbol{d}_t} p(\boldsymbol{d}_t, \boldsymbol{\ell}_j^s = c \mid \mathbf{Z}), \tag{4.17}$$

which are easily recovered from the data association probabilities stored as the com-

| | max | mean | median | min | rmse | sse | std |
|---|---|---|---|---|---|---|---|
| ML | 52.00 | 26.84 | 28.69 | 1.368 | 30.34 | 1.10e+06 | 14.14 |
| GPDA | 17.13 | 5.402 | 5.12 | **0.28** | 6.07 | 4.01e+04 | 2.77 |
| MM | **15.57** | **2.80** | **2.60** | 0.29 | **3.24** | **1.10e+04** | **1.62** |

Table 4.1: Comparison of maximum, mean, median, and minimum absolute pose error (APE) on KITTI Sequence 05 between maximum-likelihood (ML), Gaussian PDA (GPDA) and max-mixtures (MM) approaches to data association. Also provided are the root-mean-squared error (RMSE), the sum of squared errors (SSE), and the standard deviation (STD) of the APE. The best performing method in each case is shown in **bold**.

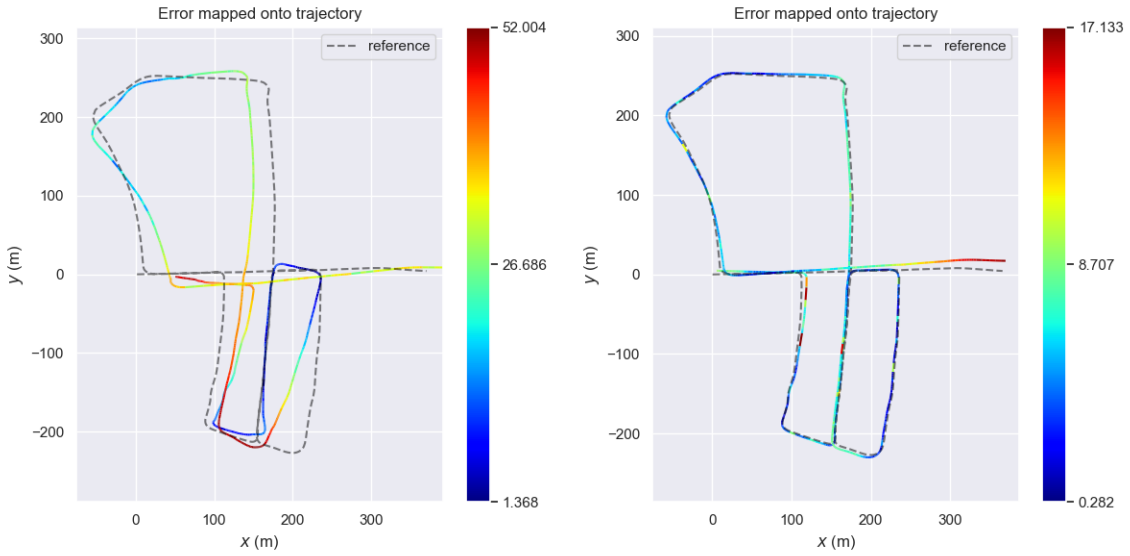
ponent weights of the max-mixture factors.

## 4.3 Experimental Results

We evaluate our approach on real stereo camera data from the KITTI dataset odometry sequence 5 [22]. In our experiments, we use the MobileNet-SSD object detector ([27, 26, 35]). We threshold the confidence of the detector at 0.8 to avoid false positive detections and use detections of cars as landmarks. We use VISO2 stereo odometry for visual odometry [23]. We estimate the range and bearing to cars as the average range and bearing to all points tracked by VISO2 that project into the bounding box for a given car detection. To solve the SLAM problem efficiently, we the implementation of iSAM2 [30] within the GTSAM [10] library. As a result, all three methods that were evaluated run at approximately 10 Hz on a single core of a 2.2 GHz Intel i7 CPU (object detections and stereo odometry were preprocessed and played back in real-time to evaluate the SLAM system). We use evo [24] for trajectory evaluation.

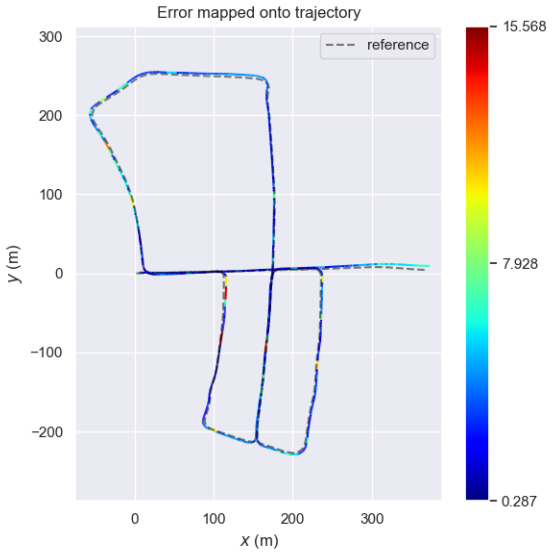
We compared maximum-likelihood data association to two-component Gaussian probabilistic data association (Gaussian PDA)[3] and our approach using a max-mixture with two components. A comparison of trajectories produced by each method is shown in Figure 4-1, which shows the estimated trajectory for each method colored by the absolute pose error (APE) at each point in the trajectory. Similarly, we pro-

[3]This is similar to a single iteration of the EM approach of Bowman et al. [5]



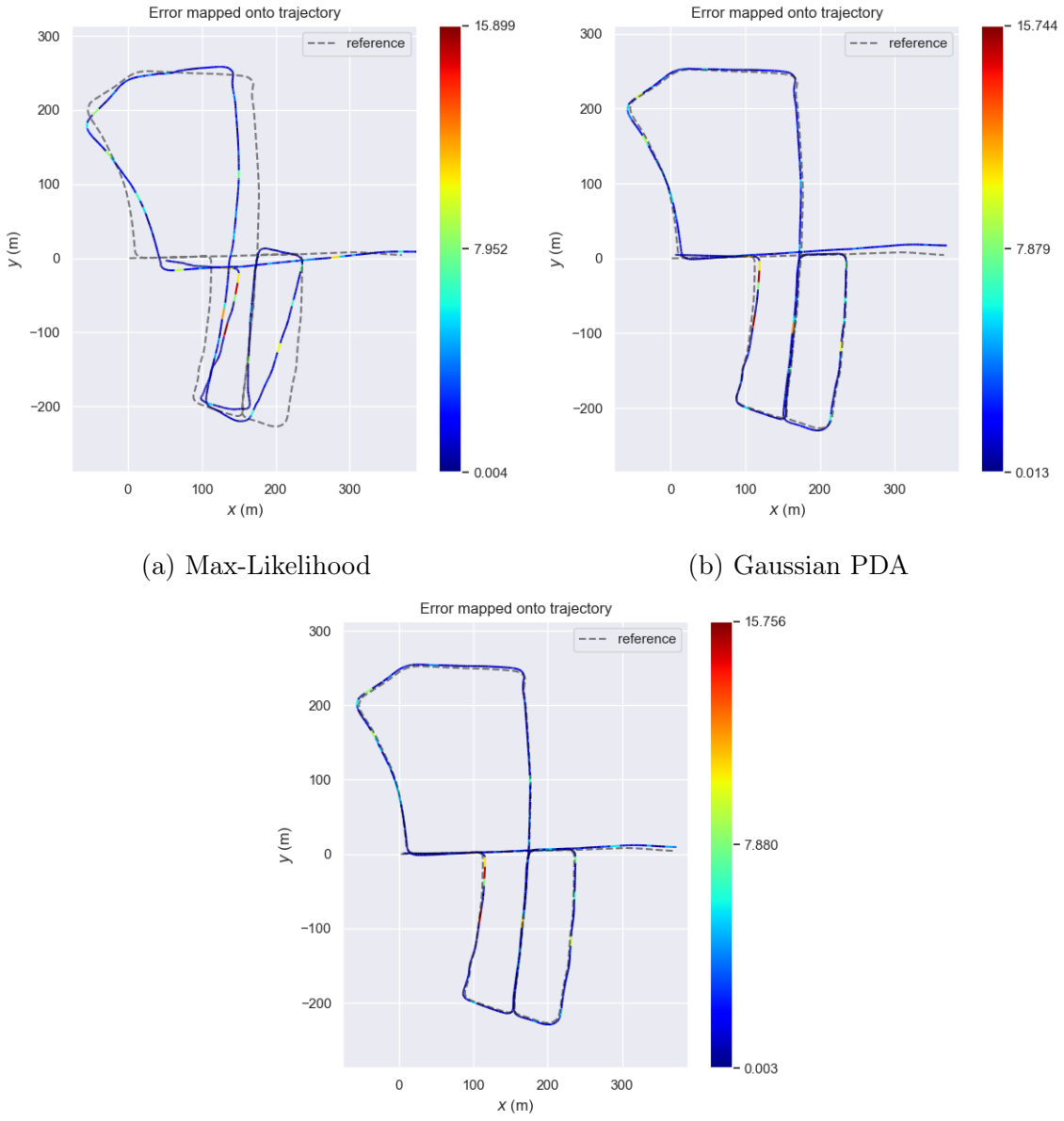

(a) Max-Likelihood

(b) Gaussian PDA




(c) Max-Mixtures

Figure 4-1: Absolute pose error (APE) mapped onto the predicted trajectory for KITTI Sequence 05. False loop closures cause the maximum-likelihood data association method to fail catastrophically, while both probabilistic methods show better performance. Note that the color is scaled uniquely to each method.

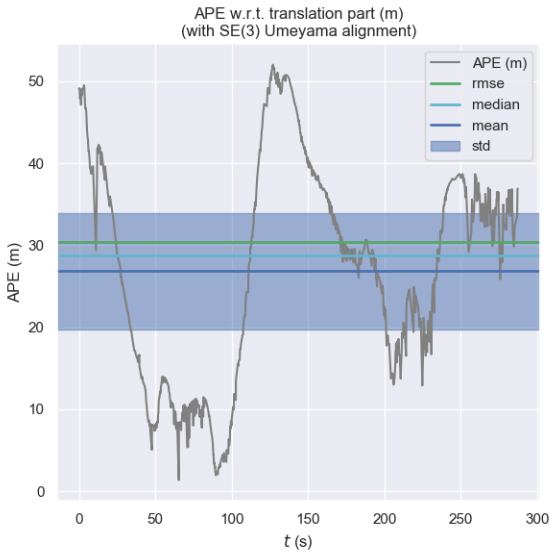



(a) Max-Likelihood

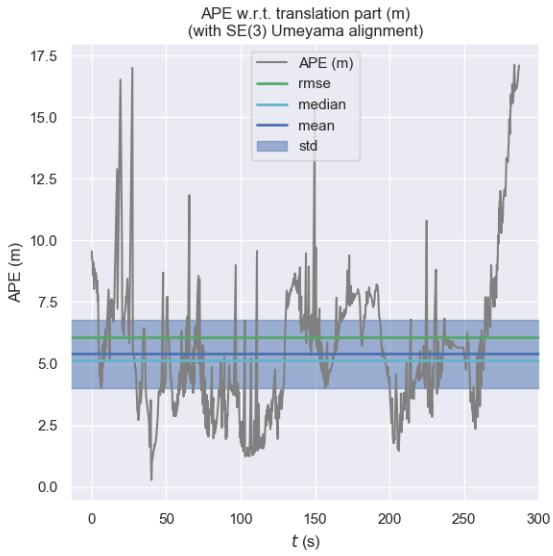
(b) Gaussian PDA

(c) Max-Mixtures

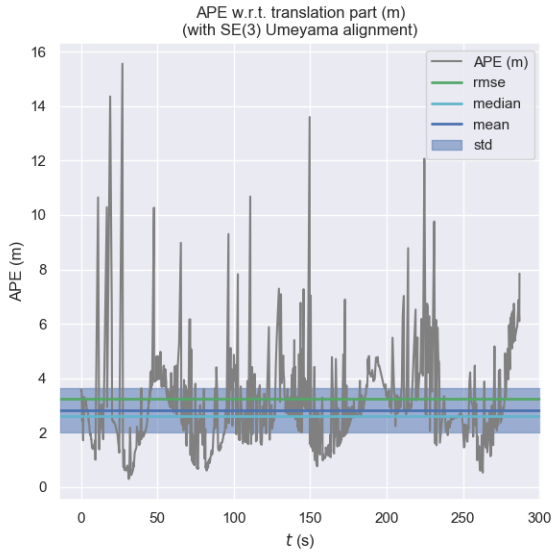
Figure 4-2: Relative pose error (RPE) mapped onto the predicted trajectory for KITTI Sequence 05. Note that the color is scaled uniquely to each method.




(a) Max-Likelihood

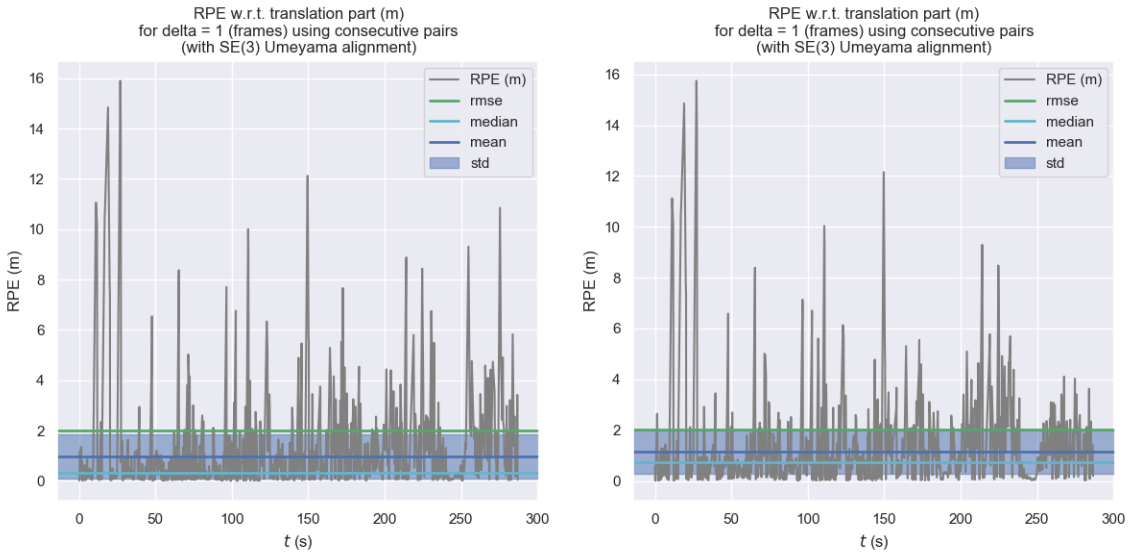



(b) Gaussian PDA



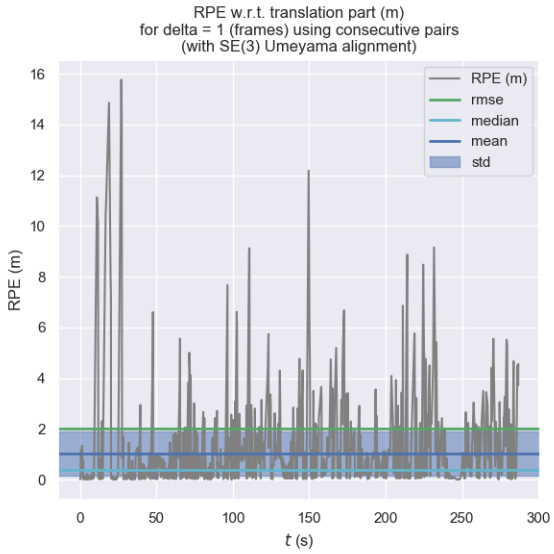

(c) Max-Mixtures

Figure 4-3: Absolute pose error (APE) over time for each method evaluated on KITTI Sequence 05. The max-mixtures approach achieved the smallest error across each metric, with Gaussian probabilistic data association performing similarly. Note the difference in $y$-axis scale across methods.




(a) Max-Likelihood

(b) Gaussian PDA




(c) Max-Mixtures

Figure 4-4: Relative pose error (RPE) over time for each method evaluated on KITTI Sequence 05. The max-mixtures approach achieved the smallest error across each metric, with Gaussian probabilistic data association performing similarly. Note the difference in $y$-axis scale across methods.

vide the vehicle trajectories colored by *relative* pose error (RPE) in Figure 4-2. The APE and RPE as a function of time for each method are shown in Figures 4-3 and 4-4, respectively. A quantitative comparison of the translation and rotation error for the methods is shown in Table 4.1. We observe that even keeping a single additional component in the max-mixture approach helps substantially in empirical performance over maximum-likelihood approaches to data association. Our method performs only slightly better than the Gaussian PDA method on this dataset, and has roughly the same minimum error, as can be seen from Table 4.1.

## 4.4 Summary

In this chapter we derived a max-mixtures approach to semantic SLAM based on max-product marginalization of data associations. The max-mixtures formulation results in a convenient nonlinear least-squares approximation to the original non-Gaussian problem. Additionally, we have demonstrated results on real data in which the proposed method empirically outperformed maximum-likelihood data association. Our results with the proposed approach were also competitive with the state-of-the-art probabilistic method (Gaussian PDA) that "averages" over all candidate associations.

# Chapter 5

# Non-Gaussian Semantic SLAM

In Chapter 4 we considered a nonlinear approximation to the non-Gaussian semantic SLAM problem. This permitted efficient approximate inference and allowed the use of many now standardized optimization tools for smoothing-based robot navigation. On the other hand, by making the max-mixture approximation to the original sum-mixture, and by linearizing measurement models, we lost the ability to represent the rich and complex uncertainties induced over our robot state by non-Gaussian measurements. In this chapter, we provide an alternative approach. Rather than converting the non-Gaussian inference problem into one of approximate nonlinear optimization under a Gaussian noise assumption, we address directly posterior inference for the non-Gaussian navigation problem. To this end, we formalize the semantic SLAM problem with uncertain data associations as one of non-Gaussian inference. To be precise, we consider full posterior inference of the form:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto \prod_i f_i(\mathbf{V}_i), \ \mathbf{V}_i \subseteq \mathbf{X}, \mathbf{L}. \qquad \text{(2.26, revisited)}$$

As discussed in Chapter 3, when we marginalize out data associations $\mathbf{D}$ to form the above inference problem, the resulting posterior is in general non-Gaussian. In this chapter, we aim to compute this posterior. We solve the inference problem approximately using nonparametric belief propagation on the Bayes tree. The specific inference procedure used is referred to as multimodal incremental smoothing and

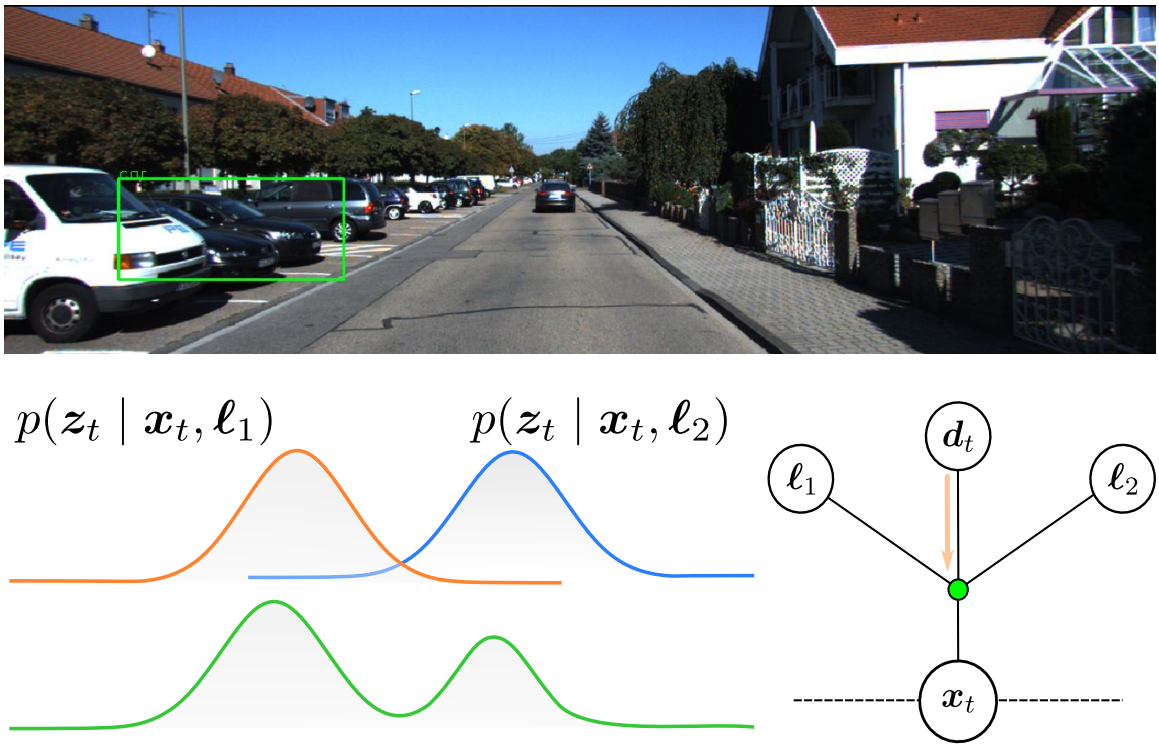



Figure 5-1: Multiple modes arise from data association ambiguity between two landmarks (1 and 2), in which the probability of an association to landmark 1 is greater than the probability of an assignment to landmark 2. *Top:* Ambiguity in an object detection results from occlusion and objects in close proximity. *Bottom-left:* Associations represented as a non-Gaussian sensor model when the data association variable is *marginalized out*. *Bottom-right*: Ambiguous measurements are incorporated into a factor graph as *multimodal semantic factors* (green). Here we depict visually the process of marginalizing the data association variable to recover a non-Gaussian factor.

mapping (mm-iSAM) [20], and we refer to the approach presented in this chapter as "multimodal semantic SLAM" [13].

## 5.1 Sum-Marginalization of Data Associations

While we previously considered "max-marginalization" of data associations, here we consider the problem of computing the sum-marginal over possible assignments to

the association variable. Our goal, in particular, is to approximate the joint belief:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) = \sum_{\mathbf{D}} p(\mathbf{X}, \mathbf{L}, \mathbf{D} \mid \mathbf{Z}) \qquad \text{(3.4, revisited)}$$
$$= \sum_{\mathbf{D}} p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}, \mathbf{Z}) p(\mathbf{D} \mid \mathbf{Z}).$$

As we have discussed, exact computation of this belief is intractable, and the result is generally a non-Gaussian posterior (for example, see Figure 5-1). The first approach we take, again, is to compute data association probabilities proactively, which greatly reduces the size of the space of data associations that need to be examined. We then solve the non-Gaussian inference problem approximately using the state-of-the-art non-Gaussian SLAM solver mm-iSAM [20].

### 5.1.1 Proactive Sum-Marginalization

We proceed with proactive computation of data association probabilities as in Chapter 4. Given a set of previous measurements $\mathbf{Z}^-$ and new measurements $\mathbf{Z}^+$ (i.e. $\mathbf{Z} = \mathbf{Z}^+ \cup \mathbf{Z}^-$ again satisfying $\mathbf{Z}^+ \cap \mathbf{Z}^- = \emptyset$). We compute the sum-marginal over associations to the new measurements, again denoted $\mathbf{D}^+$. Revisiting the joint decomposition in the previous chapter, we have:

$$p(\mathbf{X}, \mathbf{L}, \mathbf{D}^+ \mid \mathbf{Z}^+, \mathbf{Z}^-) = \frac{p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+, \mathbf{Z}^-) p(\mathbf{X}, \mathbf{L} \mid \mathbf{D}^+, \mathbf{Z}^-) p(\mathbf{D}^+ \mid \mathbf{Z}^-) p(\mathbf{Z}^-)}{p(\mathbf{Z}^+, \mathbf{Z}^-)}$$
$$\propto p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+) p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-) p(\mathbf{D}^+ \mid \mathbf{Z}^-).$$

Applying sum-marginalization to data associations, as in Equation (3.4), we obtain:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^+, \mathbf{Z}^-) \propto \sum_{\mathbf{D}^+} \left[ p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+) p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-) p(\mathbf{D}^+ \mid \mathbf{Z}^-) \right] \qquad (5.1)$$
$$= p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-) \sum_{\mathbf{D}^+} \left[ p(\mathbf{Z}^+ \mid \mathbf{X}, \mathbf{L}, \mathbf{D}^+) p(\mathbf{D}^+ \mid \mathbf{Z}^-) \right]. \qquad (5.2)$$

Exactly as in the previous chapter, $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$ is the posterior distribution over poses and landmarks after sum-marginalization of data associations to the measure-

ments $\mathbf{Z}^-$, and therefore is generally non-Gaussian. In our previous attempts to convert this problem into one of nonlinear least-squares, we made a number of approximations here. In particular, when computing the max-marginal, we replaced the joint belief with its max-marginal form, then approximated further by selecting only a single set of data associations when computing data association probabilities $p(\mathbf{D}^+ \mid \mathbf{Z}^-)$. Here, we consider approximate non-Gaussian inference of the posterior, allowing us to directly address problems of this form.

In the recurring example of a factor graph with two vehicles poses linked by an odometry constraint, each measuring one of two candidate landmarks, each with a prior (see Figures 3-1 and 3-3), we earlier wrote the posterior as:

$$p(\mathbf{X}, \mathbf{L}, \mathbf{D} \mid \mathbf{Z}) \propto \psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2)\rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_1)\rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_2)\phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2),$$

where here $\mathbf{D} = \{d_1, d_2\}$. In this chapter, we allow the distribution corresponding to any or all of these factors to be an arbitrary non-Gaussian distribution. That is, we make no specific distributional assumptions whatsoever (in contrast to the max-mixtures approach). Exact marginalization of the data association variables corresponds to the following computation:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto \sum_{\mathbf{D} \in \mathbb{D}} \psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2)\rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_1)\rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_2)\phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2). \quad (5.3)$$

Assuming that data associations are independent (conditioned on previous measurements) we push the summation into the individual terms. This allows us to obtain the following result:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto \psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2)\phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2) \sum_{d_1 \in \mathcal{D}_1} \rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_1) \sum_{d_2 \in \mathcal{D}_2} \rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_2). \quad (5.4)$$

Since we have relaxed the assumption of Gaussianity, we can simply rewrite this

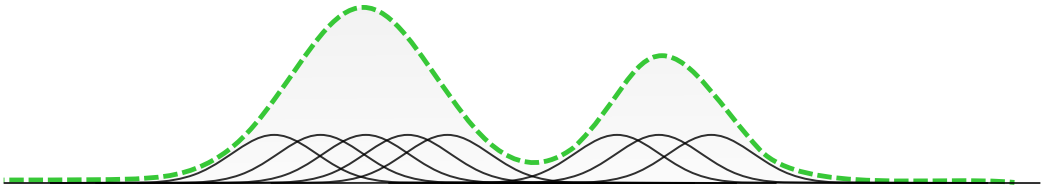


Figure 5-2: The representation used by nonparametric belief propagation consists of a mixture of evenly-weighted Gaussian kernels. In order to approximate a distribution **(green)**, nonparametric belief propagation (and mm-iSAM, consequently) uses a fixed number of evenly-weighted kernels **(black)**.

expression by defining the following new non-Gaussian factors:

$$f_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \triangleq \sum_{d_1 \in \mathcal{D}_1} \rho_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_1) \tag{5.5}$$

$$f_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2) \triangleq \sum_{d_2 \in \mathcal{D}_2} \rho_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2, d_2), \tag{5.6}$$

which gives the new joint belief:

$$p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}) \propto \psi_1(\boldsymbol{x}_1, \boldsymbol{x}_2)\phi_1(\boldsymbol{x}_1)\varphi_1(\boldsymbol{\ell}_1)\varphi_2(\boldsymbol{\ell}_2)f_1(\boldsymbol{x}_1, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2)f_2(\boldsymbol{x}_2, \boldsymbol{\ell}_1, \boldsymbol{\ell}_2). \tag{5.7}$$

Here we refer back to the sum-product algorithm (Algorithm 1) for details about the above marginalization. Similarly, this corresponds to the elimination of the data association variable as displayed in Figure 3-3.

The number of terms in the product of these summations grows quite quickly. Hence, we do not explicitly compute distributions of the above form. Rather, we resort to approximate Bayesian inference.

## 5.2 Multimodal iSAM

We use multimodal iSAM (mm-iSAM) [20] to compute the posterior over poses and landmarks given a non-Gaussian factor graph, such as those often obtained after marginalization of data association variables. To accommodate non-Gaussian variables in the factor graph, multimodal iSAM makes use of nonparametric belief propagation [52]. Nonparametric belief propagation approximates the belief over all con-

tinuous state variables absent the assumption of Gaussianity using a combination of Gibbs sampling and kernel density estimation. That is, for a random variable $X$, we approximate the marginal over $X$ as

$$\hat{p}(X) = \sum_{n=1}^{N} w^{[n]} \mathcal{N}\left(x^{[n]}, \Sigma^{[n]}\right), \tag{5.8}$$

where $\mathcal{N}$ is a multivariate Gaussian kernel, each kernel is centered at a sample $x^{[n]}$, $w^{[n]}$ is the weight associated with the $n$-th kernel, and $\Sigma^{[n]}$ is the associated Gaussian kernel bandwidth, determined using leave-one-out cross-validation. The weights $w^{[n]}$ are chosen uniformly such that the resulting sum is a valid probability density function. An illustrative example of the belief representation used by nonparametric belief propagation (and consequently mm-iSAM) is shown in Figure 5-2.

### 5.2.1 Computational Complexity

A beneficial aspect of the functional approximation of marginals is that we no longer need to explicitly represent the potentially many modes in the posterior. This *implicit* representation decouples the complexity of inference from the number of hypotheses, as the computation involved in the approximation of the marginals depends only on a fixed number of samples. Due to proactive computation of data associations, examining the space of candidate hypotheses has the same complexity as the max-mixtures approach (generally roughly linear in the number of measurements). Similar to the max-mixtures approach, the process of inference is agnostic to the number of hypotheses. In the case of max-mixtures, the optimization corresponds to nonlinear least-squares, and the representation is essentially no different from any other nonlinear measurement function from the perspective of the optimizer. In the multimodal (non-Gaussian) SLAM case, the complexity of inference is no different than it would be for any other arbitrary non-Gaussian distribution.

Both the complexity and accuracy of inference depend critically on the number of samples used for inference, which is a hyperparameter selected by the practitioner (determined *a priori*) and controls the fidelity of the non-Gaussian approximation.

The result is that modes with very low probability are unlikely to be represented in the approximate marginal density. However, we do not explicitly prune these modes, and since they still exist in the factor graph, modes which later become more probable can be recovered.

Of practical note, the nonparametric inference method used by multimodal iSAM is in its present form computationally intensive relative to mature nonlinear least-squares solvers. At the time the experiments in this thesis were performed, a single solve involving around 300 vehicle poses and less than 100 landmarks required around 10 seconds of computation time. Re-solving several times over the course of a trajectory often required a cumulative computation time of several minutes. There are a number of optimizations that could be made to the implementation of the nonparametric inference procedure, which we discuss in Chapter 6, but such optimizations are outside the scope of the work in this thesis.

## 5.3 Multimodal Semantic Factors

Earlier in this chapter, we described how we could arrive at non-Gaussian factors describing sum-marginalization of data associations. We now aim to explicitly describe these factors. As in Chapter 4, marginalization of data associations critically depends on computation of the distribution over new data associations given all previous measurements $p(\mathbf{D}^+ \mid \mathbf{Z}^-)$.

To incorporate the non-Gaussian factors resulting from ambiguous data associations into the factor graph, we use multimodal factors. In particular, since we consider measurements with semantic properties, we consider multimodal semantic factors, which introduce constraints between a pose and potentially many landmarks. As we have throughout this thesis, we assume a factorized semantic measurement model $p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\ell}) = p(\boldsymbol{z}^s \mid \boldsymbol{\ell}^s)p(\boldsymbol{z}^r \mid \boldsymbol{x}, \boldsymbol{\ell})p(\boldsymbol{z}^b \mid \boldsymbol{x}, \boldsymbol{\ell})$ consisting of the class estimate $\boldsymbol{z}^s$ from an object detector, the estimated range to the object $\boldsymbol{z}^r$, and the estimated bearing to the object $\boldsymbol{z}^b$. The distribution $p(\boldsymbol{z}^s \mid \boldsymbol{\ell}^s)$ corresponds to the confusion matrix for the classifier, learned offline, while $p(\boldsymbol{z}^r \mid \boldsymbol{x}, \boldsymbol{\ell})$ and $p(\boldsymbol{z}^b \mid \boldsymbol{x}, \boldsymbol{\ell})$ are each

assumed Gaussian with means $\boldsymbol{z}^r$ and $\boldsymbol{z}^b$ and variances $\sigma_r^2$ and $\sigma_b^2$, respectively. We determine the latter terms by considering the range and bearing to the set of 3D points estimated by a stereo vision system which project into the bounding box for the object detection corresponding to measurement $\boldsymbol{z}$.

### 5.3.1 Monte Carlo Approximation of Association Probabilities

At each time step $t$, we update the factor graph solution in order to obtain the belief $p(\mathbf{X}, \mathbf{L} \mid \mathbf{Z}^-)$, which provides marginals for all poses $\mathbf{X} \triangleq \boldsymbol{x}_{1:t}$ and known landmarks. Given the semantic measurement model, we compute the probability of an association as the total posterior probability of all associations at time $t$ of measurement $k$ to landmark $j$ given the measurements. That is, we consider *joint-compatibility* for the measurements taken at time $t$, as opposed to *individual-compatibility*, which we used in the case of max-mixtures[1]. Let $\mathcal{D}_t$ denote the set of all possible associations of measurements at time $t$ to known landmarks. Similarly, define $\mathbb{D}_t^k(j) \triangleq \{\boldsymbol{d}_t \in \mathcal{D}_t \mid d_t^k = j\}$, the set of all possible sets of data associations at time $t$ in which measurement $k$ is associated to landmark $j$. Assuming a uniform prior on data associations, we then have:

$$p(d_t^k = j \mid \mathbf{Z}^-) \propto \sum_{\boldsymbol{d}_t \in \mathbb{D}_t^k(j)} \prod_{i=1}^{K_t} p(\boldsymbol{z}_t^i \mid \boldsymbol{d}_t, \mathbf{Z}^-), \tag{5.9}$$

Just as in the previous chapter, we compute the likelihood of each measurement $\boldsymbol{z}_t^k$ given its association $d_t^k$ by marginalizing out the pose estimate at $\boldsymbol{x}_t$ and the landmark

[1]Neither of our proposed methods is restricted to joint-compatibility or individual compatibility.

position and class, i.e. for arbitrary $\boldsymbol{z}_t$:

$$\begin{aligned} p(\boldsymbol{z}_t^k \mid \boldsymbol{d}_t, \mathbf{Z}^-) &= p(\boldsymbol{z}_t^k \mid d_t^k = j, \mathbf{Z}^-) \\ &= p(\boldsymbol{z}_t^s \mid d_t^k = j, \mathbf{Z}^-) p(\boldsymbol{z}_t^r, \boldsymbol{z}_t^b \mid d_t^k = j, \mathbf{Z}^-) \\ &= \left[ \sum_c p(\boldsymbol{z}_t^s \mid \boldsymbol{\ell}_j^s = c) p(\boldsymbol{\ell}_j^s = c \mid \mathbf{Z}^-) \right] \times \dots \\ &\dots \left[ \iint p(\boldsymbol{z}_t^r, \boldsymbol{z}_t^b \mid d_t^k = j, \boldsymbol{x}_t, \boldsymbol{\ell}_j) p(\boldsymbol{x}_t, \boldsymbol{\ell}_j \mid \mathbf{Z}^-) d\boldsymbol{x}_t d\boldsymbol{\ell}_j \right]. \end{aligned}$$

Above we have dropped the superscript $k$ for the measurement in order to make explicit the dependence on the range, bearing, and semantic components. Furthermore, here we take $j$ as the particular association for measurement $\boldsymbol{z}_t^k$, i.e. it does not correspond to the $j$ in Equation (5.9). Similarly, here we take $k$ as an arbitrary index into the measurements at time $t$, and $k$ does not correspond to $k$ in Equation (5.9). Rather, $k$ is replaced with each $i$ as we compute the product over all measurement probabilities given the set of data associations and history of measurements at the current time $t$.

While in Chapter 4 we approximated this integral by considering a single component of the max-marginal, in a non-Gaussian framework, we can readily compute this term by Monte Carlo approximation using a set of pose samples $\boldsymbol{x}_t^{[n]}, n = 1, \dots, N$:

$$\begin{aligned} p(\boldsymbol{z}_t^k \mid \mathbf{D}_t, \mathbf{Z}^-) \approx & \left[ \sum_c p(\boldsymbol{z}_t^s \mid \boldsymbol{\ell}_j^s = c) p(\boldsymbol{\ell}_j^s = c \mid \mathbf{Z}^-) \right] \times \dots \\ & \dots \sum_{n=1}^{N} \int p(\boldsymbol{z}_t^r, \boldsymbol{z}_t^b \mid \boldsymbol{x}_t^{[n]}, \boldsymbol{\ell}_{d_t^k}, d_t^k) p(\boldsymbol{x}_t^{[n]} \mid \mathbf{Z}^-) p(\boldsymbol{\ell}_{d_t^k} \mid \mathbf{Z}^-) d\boldsymbol{\ell}_{d_t^k}, \end{aligned} \tag{5.10}$$

where we have replaced the integral over the pose distribution by a sampled approximation. For data association computation, we adopt a maximum likelihood sensor model to simplify the integral over the landmark position. We find this works well, empirically, when the sensor model is Gaussian, but non-Gaussian sensor models can be accommodated by making a sample-based approximation, for example. An illustrative example of the computation in Equation (5.10) is shown in Figure 5-4.

### 5.3.2 Constructing Multimodal Semantic Factors

Given $p(d_t^k = j \mid \mathbf{Z}^-)$ for all landmarks $j$ in a set $\mathcal{H} \subseteq 1, \ldots, M$ of candidate landmark *hypotheses*, a multimodal semantic factor for each measurement $k$ links a pose $\boldsymbol{x}_t$ and each candidate in $\mathcal{H}$:

$$f(\boldsymbol{x}_t, \boldsymbol{\ell}_{\mathcal{H}}) = \sum_{j \in \mathcal{H}} p(\boldsymbol{z}_t^k \mid \boldsymbol{x}_t, \boldsymbol{\ell}_j) p(d_t^k = j \mid \mathbf{Z}^-), \tag{5.11}$$

which for a Gaussian measurement model is a weighted sum of Gaussians.

Finally, MAP estimates for each landmark class, assuming a uniform prior, can be computed in the same manner as in Chapter 4:

$$\boldsymbol{\ell}_j^c = \underset{c}{\operatorname{argmax}} \prod_t \sum_{\boldsymbol{d}_t} p(\boldsymbol{d}_t, \boldsymbol{\ell}_j^c = c \mid \mathbf{Z}), \tag{4.17, revisited}$$

which are obtained by maximizing over the probabilities determined using Equation 5.10 with respect to the landmark classes, rather than marginalizing them out, and instead marginalizing out data associations.

## 5.4 Experimental Results

Experiments with mm-iSAM were implemented in the Julia programming language using the Caesar.jl library[2]. We demonstrate the proposed approach both in simulation, with a hallway environment, and using real data from the KITTI dataset [22, 21]. All experiments were run offline using 10 cores of a 2.2 GHz i7 CPU and factor graph computation time was roughly identical across the three methods (approximately 1 minute for simulated data and 3 minutes for the KITTI dataset). In both tests, we compared our method, multimodal semantic SLAM (MMSS) with maximum-likelihood (ML) data association and Gaussian probabilistic data (Gaussian PDA). The ML method selects the maximum-likelihood association considering all measurements in a keyframe. We implement the Gaussian PDA method using

[2] `https://github.com/JuliaRobotics/Caesar.jl`

Gaussian factors with variance inversely weighted by data association probabilities[3].

In practice, new landmarks are determined using a threshold on their likelihood given each known landmark (similar to a Mahalanobis distance threshold in the Gaussian case) and we compute data association probabilities for each candidate landmark within a conservative range of the estimated pose at time $t$ (this determines the set $\mathcal{H}$ in Equation 5.11).

### 5.4.1 Simulated Data

Our simulated navigation experiments consist of a two-dimensional hallway environment with landmarks of two classes. The robot in this simulation makes noisy measurements to each landmark within its limited field of view (120° up to 3.5 m), and each range measurement has an associated distribution over class probabilities. We model semantic measurements as samples from a categorical model having a confusion matrix with 90% accuracy for all landmark classes. Range and bearing measurements were corrupted with zero-mean Gaussian noise with variance 0.01. We also simulate an odometry model corrupted by Gaussian noise with diagonal covariance $\Lambda_t$, which we vary in our experiments.

In Figures 5-3a-c, trajectories and landmark estimates from each method are compared qualitatively for a simulated run with $\Lambda_t = \text{diag}(0.01; 0.01; 0.001)$. In this example, we find that ML data association fails in the presence of substantial perceptual aliasing. Both Gaussian PDA and our method are more robust to errors in data association, but we find that ours is the only method that accurately closes the loop after executing the full trajectory. In Figure 5-3d, we show the average trajectory error for the three methods, plotted against $tr(\Lambda_t)$. Error for our approach increases the least as the odometry becomes more noisy, suggesting improved robustness to odometry uncertainty.

[3]Our implementation of the Gaussian PDA method uses the approximate marginal likelihood of each observation to compute data association probabilities, rather than a point estimate of poses and landmarks; thus, it can be viewed as an extension of the EM formulation in [5] from maximum-likelihood estimation to MAP estimation.

### 5.4.2 Real Data

We evaluated the three approaches for a navigation using a stereo camera with data from KITTI odometry sequence 5 [22]. Odometry is provided by VISO2 stereo odometry [23], and probabilistic data associations with objects provide loop closures. We sample keyframes at 1 Hz and objects are detected in the left camera image using the MobileNet-SSD neural network [27] (with the single-shot detector (SSD) and MobileNets proposed respectively in [35] and [26]) trained on the PASCAL Visual Object Categories (VOC) dataset [17]. We accept measurements for which the neural network reports a confidence greater than 0.8. Semantic measurements are produced in the KITTI dataset by detections of cars and are represented by the average range and bearing to all 3D points that project into the detection bounding box. We assume that the stereo pair has fixed height and is constrained in pitch and roll, so the resulting estimation procedure is carried out with respect to the vehicle translation along the ground plane and yaw.

Figure 5-5 shows estimated trajectories and landmark positions for each method on KITTI sequence 5, and corresponding average translation and rotation errors can be found in Table 5.1. As a result of perceptual aliasing due to long rows of parked cars, maximum-likelihood associations cause a number of incorrect loop closures that are hard to recover from. Gaussian PDA makes "soft" measurements in these cases, and produces a much better solution. By representing the full posterior, however, our method obtains a more accurate solution, recovering from the uncertainty in growth in the largest loop. We additionally mark a pose near this loop closure in Figure 5-5c and display the contour plot of its distribution in Figure 5-6, which shows that odometry uncertainty coupled with data association ambiguity results in a non-Gaussian posterior. A supplemental video provides visualization of the object detections and estimated vehicle trajectory using our approach on the KITTI dataset[4].

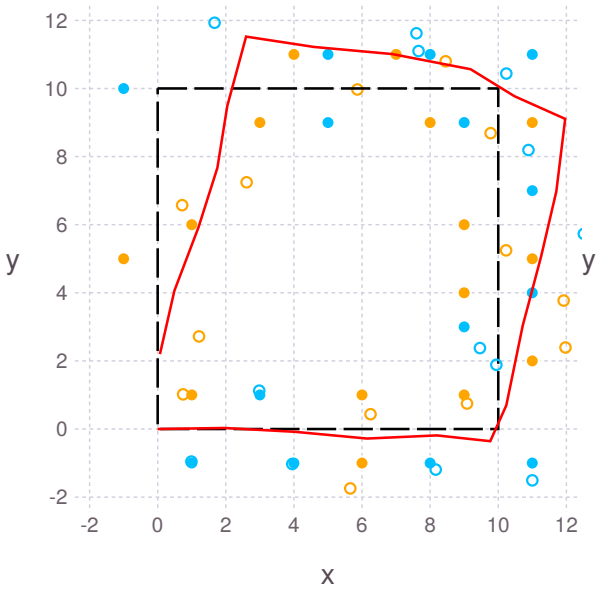
[4] `https://youtu.be/9hEonD8KDrs`

| Method | Avg. Trans. Error (m) | Avg. Rot. Error (rad) |
|---|---|---|
| ML | 20.427 | 0.0810 |
| GPDA | 8.814 | 0.0446 |
| MMSS (Ours) | **5.718** | **0.0255** |

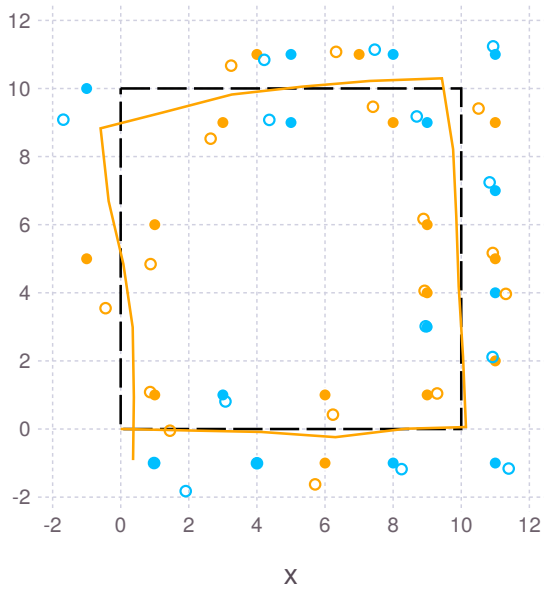
Table 5.1: Comparison of translation and rotation error on KITTI sequence 5 for the different methods tested.

## 5.5 Summary

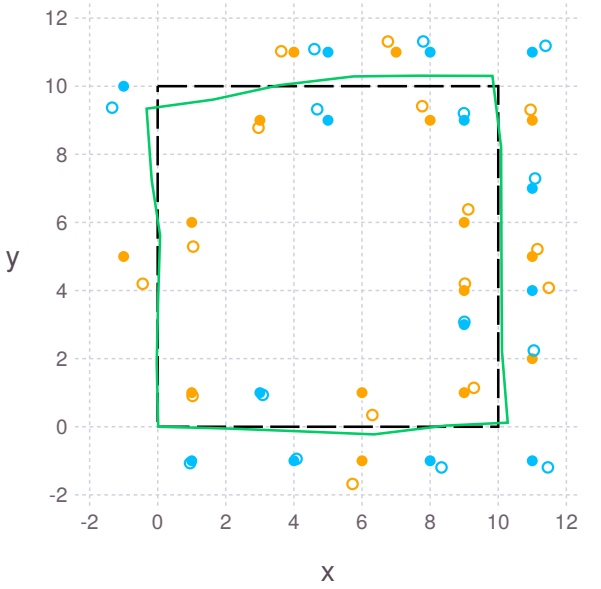
In this chapter, we proposed a solution to semantic SLAM with unknown data associations that implicitly represents multiple association hypotheses as a multimodal sensor model. This formulation leads to a non-Gaussian SLAM problem, which we solve using mm-iSAM [20]. Constructing our approach with non-Gaussian inference in mind, in the process we developed a method of performing semantic SLAM with unknown data association in situations where measurements themselves are non-Gaussian (e.g. when using acoustic sensing), and consequently can operate with many sensors with characteristics that are difficult or impossible to model adequately in traditional SLAM frameworks. We validated our approach on a simulated navigation task under variety of odometry noise characteristics, as well as on data from the KITTI dataset. In addition to representing non-Gaussian belief over poses and landmarks, our multimodal semantic SLAM approach showed improved robustness to odometry noise and perceptual aliasing as compared with other methods.



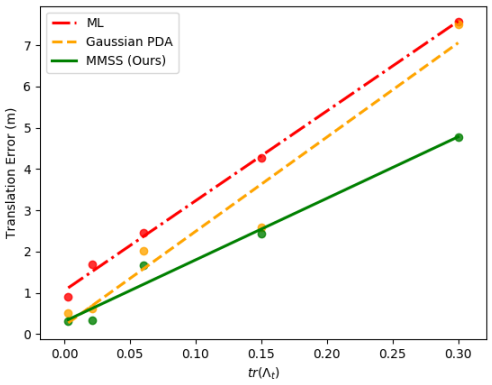

(a) Maximum-Likelihood




(b) Gaussian PDA




(c) MMSS (Ours)




(d) Avg. translation error vs. $tr(\Lambda_t)$

Figure 5-3: (a-c): Comparison of trajectories estimated using each approach in a simulated hallway environment. Ground-truth trajectories are shown as dashed black lines. Ground-truth landmarks are shown as circles and colored by semantic class. Landmark position estimates from each method are shown as rings and colored similarly by class. (d) Comparison of translation error on simulated navigation tasks under five odometry noise models, $\Lambda_t$, with best fit line for each method.

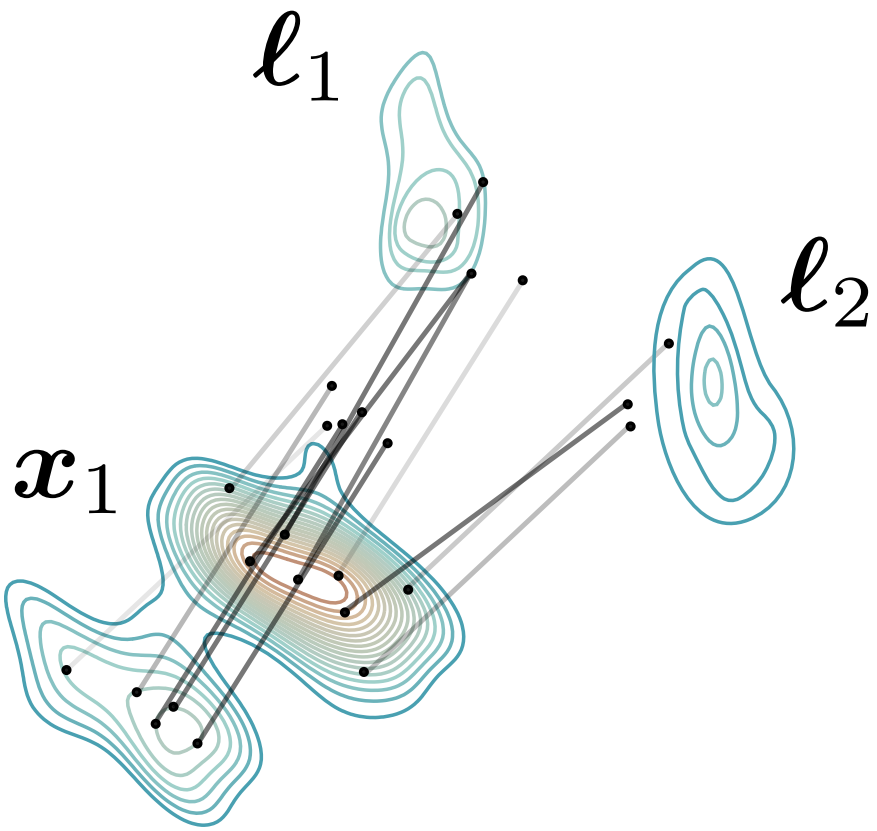
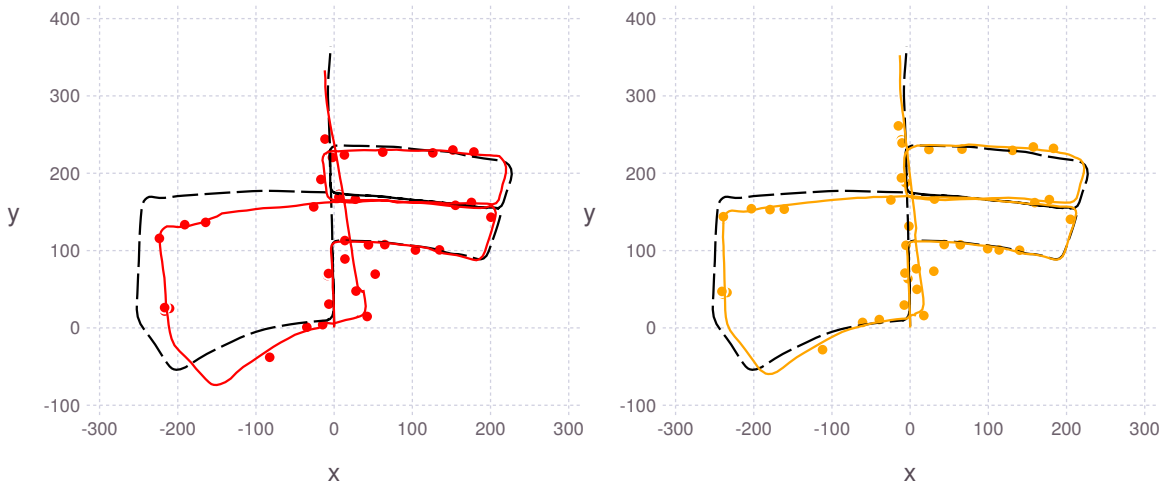
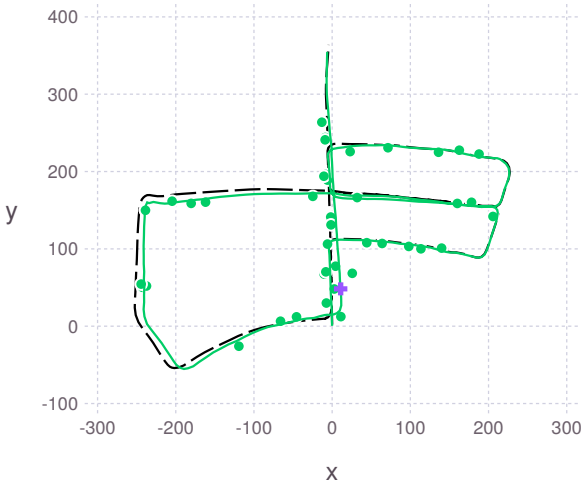



Figure 5-4: Illustrative example of approximating data association probabilities as in Equation (5.10). Each contour plot describes the marginal distribution for one of the three relevant variables $\boldsymbol{x}_1$, $\boldsymbol{\ell}_1$, or $\boldsymbol{\ell}_2$. We use a sample-based approximation to compute data association probabilities, wherein we draw pose samples $\boldsymbol{x}_1^{[n]}$ from the marginal distribution over poses $p(\boldsymbol{x}_1 \mid \mathbf{Z}^-)$. We then approximate Gaussian measurement models using the maximum-likelihood estimate, i.e. we take $p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\ell})$ as $\delta(h(\boldsymbol{x}, \boldsymbol{\ell}); \boldsymbol{z})$ where $\delta(\cdot)$ here is the Dirac delta function equal to 1 only where $h(\boldsymbol{x}, \boldsymbol{\ell}) = \boldsymbol{z}$, and zero elsewhere. This produces the terminal points of each line (corresponding to range-bearing measurements) in the above figure. The opacity of each line represents the probability of the pose $\boldsymbol{x}_1^{[n]}$ from which it originates. The result is a set of weighted Dirac functions corresponding to the locations where the most-likely measurement would "land" in space from the pose sample $\boldsymbol{x}_1^{[n]}$. The product of these functions with each landmark marginal represents the result of the approximate convolution in Equation (5.10).




(a) Maximum-Likelihood

(b) Gaussian PDA




(c) MMSS (Ours)

Figure 5-5: Comparison of trajectories (lines) and landmark position estimates (points) for each method applied to KITTI sequence 5. Ground truth trajectory is plotted as a black dashed line. The contour plot for the pose marked with a purple cross in (c) is shown in Figure 5-6.

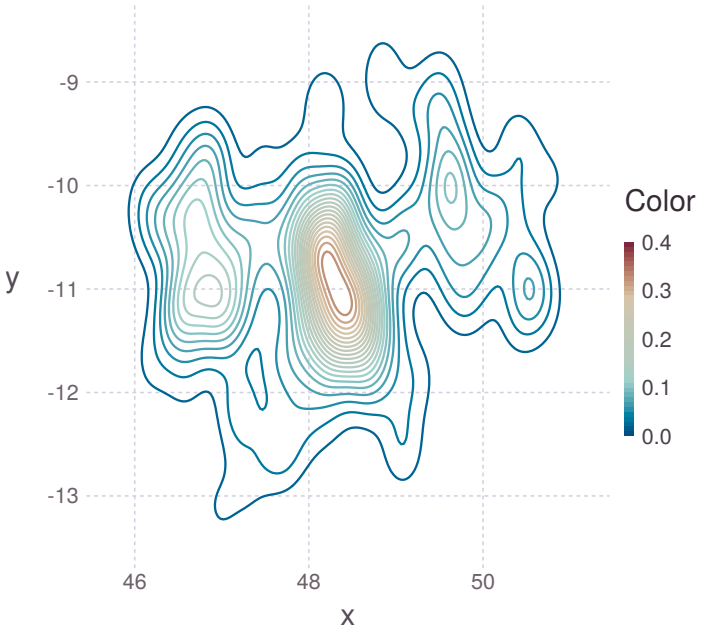



Figure 5-6: Contour plot for the marginal distribution of the marked pose in Figure 5-5c. Multimodality is induced by odometry uncertainty and data association ambiguity.

# Chapter 6

# Discussion and Conclusion

## 6.1 Our Contributions

In this thesis we developed a formulation of semantic SLAM with unknown data association wherein we use variable elimination to remove the data association variables from the inference process. We leverage both the *max-product* and *sum-product* variable elimination algorithms for performing marginalization of data association variables. Considering each of these algorithms led to the development of two novel approaches for semantic SLAM, "max-mixtures semantic SLAM" and the non-Gaussian method "multimodal semantic SLAM". We have shown through experiments on real and simulated data that the proposed methods are competitive with the state-of-the-art methods for semantic data association.

### 6.1.1 Max-Mixtures Semantic SLAM

The max-product approach we developed has a strong correspondence with the well-known "max-mixtures" approach for robust SLAM (hence our reference to the method as "max-mixtures semantic SLAM"), and allows us to perform approximate MAP inference under the assumption of nonlinear Gaussian measurement models using state-of-the-art nonlinear optimization method for SLAM like iSAM2 [30].

The max-mixtures approach required us to make several approximations in the

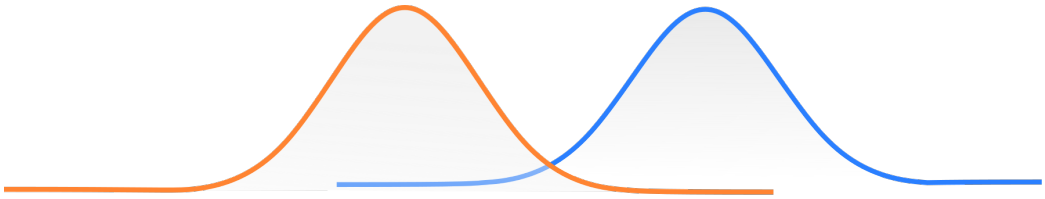
computation of data associations. Most critically, we choose an assignment to all other data associations when computing new data association probabilities. Beyond that, however, the assumption of nonlinear Gaussian measurement models is quite limiting in the context of sensors like sonar, or measurements which are "undetermined" e.g. range-only or bearing-only measurements.

### 6.1.2 Multimodal (Non-Gaussian) Semantic SLAM

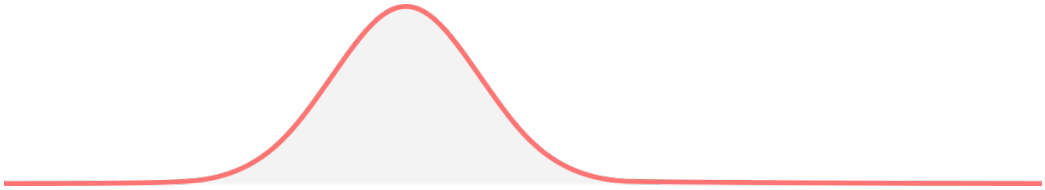
The second approach we consider, based on the sum-product marginalization of data associations, corresponds to a problem of non-Gaussian Bayesian inference. Using the state-of-the-art non-Gaussian SLAM solver multimodal-iSAM (mm-iSAM) [20], we are able to perform approximate inference on factor graphs with arbitrary non-Gaussian factors. This allows us to relax the assumptions of the max-mixture approach and infer complex, multimodal posterior distributions with nontrivial dependence on the history of data associations. Furthermore, this allowed us to obtain a more comprehensive approximation of the probabilities of data associations as compared to the computation we used for the max-mixtures approach. In particular, the non-Gaussian approach leverages a sample-based approximation that considers the full, non-Gaussian posterior marginals over poses and landmarks, rather than assuming a particular set of data associations in order to force the resulting integral to be over a product of Gaussian terms.

## 6.2 Comparison of Representations

In Figure 6-1, we show an illustrative comparison of all of the data association methods described throughout this thesis. In particular, we compare the representations of maximum-likelihood data association, the single iteration expectation-maximization (EM) approach, referred to in this thesis as Gaussian probabilistic data association (PDA), the max-mixtures approach of Chapter 4, and the multimodal (non-Gaussian) approach of Chapter 5. In this figure, we assume the probability of the left (orange) hypothesis is greater than that of the right (blue) hypothesis.



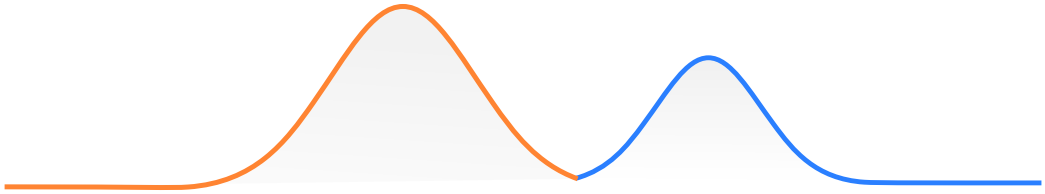
(a) Multi-hypothesis ambiguity for two associations.



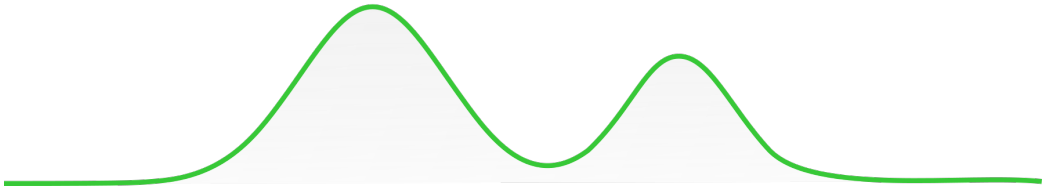
(b) Maximum-likelihood data association approach.



(c) Single iteration EM approach (Gaussian Probabilistic Data Association.)



(d) Max-mixtures approach (used in Max-Mixtures Semantic SLAM)



(e) Non-Gaussian approach (used in Multimodal Semantic SLAM).

Figure 6-1: Illustrative comparison of the maximum-likelihood, Gaussian probabilistic data association, max-mixture, and sum-mixture representations for data association ambiguity. Here we assume the probability of the leftmost hypothesis **(orange)** is greater than that of the rightmost hypothesis **(blue)**.

The maximum-likelihood approach selects the mode with the larger probability (the orange mode), and discards the alternative mode. Since each hypothesis corresponds to a Gaussian solution (conditioned on the data association), the result is a single Gaussian mode.

The Gaussian probabilistic data association approach also obtains a unimodal Gaussian solution. However, with only a single iteration of expectation-maximization, this approach corresponds to computing a geometric mean over possible solutions given each data association. The resulting unimodal Gaussian solution is the "average" over possible associations, and consequently sits "between" the candidate modes, closer to the mode with higher probability (left).

The max-mixtures approach selects the most probable data association at each point in the measurement domain. The result is that the association "switches" from one hypothesis (left) to the other (right) when the probability of the latter hypothesis (blue) becomes greater than that of the former hypothesis (orange). At any given point, the solution "looks" Gaussian locally, and is amenable to nonlinear least-squares optimization methods if we seek the most probable estimates of poses and landmarks. Note that characterization of uncertainty in this framework is quite difficult, as linearization of the covariance implicitly requires selecting a set of data associations (as we do when computing data association probabilities in Chapter 4).

Lastly, the multimodal approach computes the proper sum-marginal over candidate data associations. In the Gaussian case, this sum results in a sum-of-Gaussians weighted by the individual association probabilities. The resulting mixture is not in a form that can be addressed by standard nonlinear least-squares methods for SLAM. Consequently, we resort to new tools for non-Gaussian inference to infer the posterior distribution over poses and landmarks. Inference in this framework is necessarily approximate, but allows for a higher fidelity characterization of uncertainty than the max-mixtures approach, which required implicitly selecting a set of data associations when performing linearization. Furthermore, the individual measurements need not be Gaussian for this approach. Given an arbitrary (non-Gaussian) measurement model, the resulting sum-mixture would simply be a weighted combination of

measurement models with one component corresponding to each possible landmark.

## 6.3 Limitations of the Proposed Approaches

Despite the promising initial results of the proposed approaches, there remain several limitations of these methods in their current form that ought to be addressed. In particular, proactively computing data association weights and never revising them allows us to easily construct situations where the correct association may have very low probability. This can happen if, for example, there is substantial drift in odometry over a large loop and too much ambiguity in possible data associations due to accumulated uncertainty. The ability to revise data associations has potential to greatly improve both of these methods. This could be done in the non-Gaussian case, for example, by introducing the discrete data association variables into the Bayes tree and performing Gibbs sampling when estimating the joint probability for a clique.

Even more commonly, we may have poorly modeled the covariance (or more broadly the error distribution) of one of our measurements, such as the stereo odometry, giving an incorrect characterization of measurement uncertainty. More work is needed on accurate characterization of uncertainty in black-box systems like those often used for visual odometry. One benefit, however, of the non-Gaussian approach is that we can make use of arbitrary, potentially non-Gaussian, empirical noise models computed offline using validation data. This is especially important in the characterization of noise for sensors like sonars, but also for characterizing the noise distribution of black-box models like visual odometry subsystems that may have environment dependent noise.

Finally, throughout this thesis, we made the assumption that our detector produces no "false positives", i.e. detections that arise without correspondence to any object at all. This assumption allowed us to focus on errors due to misclassification or ambiguity in the location of detected objects. In reality, is incredibly difficult to achieve zero false positives for any detector. For practical purposes, we reduce the potential for false positives in this work by setting a high detection confidence thresh-

old. This causes us to miss a large number of detections that otherwise could provide informative loop closures in order to avoid the detrimental effects of detector false positives. Better characterization of the errors made by object detectors is vital for their use in navigation systems.

## 6.4 Future Work

Both of the methods we presented could be very promising when combined with methods for place recognition. In particular, place recognition could help recover in situations where perhaps we cannot reliably make a data association due to odometry drift. Another idea that we could benefit from is *uniqueness* in data association. Specifically, by maintaining a feature descriptor of relevant objects in the scene, we can enable more precise data association based not just on the class of the object, but also the unique features of that specific object.

While our methods deal with uncertain associations, like many previous efforts, we rely on hard decisions about whether or not to add landmarks. Representing this uncertainty is an important step toward more tightly coupling the data association and SLAM problems, but computation of association probabilities may become expensive. Dirichlet process priors on associations, as in [38] provide one avenue for future work, while the approximate matrix permanent methods of [2] may help address computational complexity.

Besides the ability to make use of arbitrary non-Gaussian measurement models, another benefit of the multimodal semantic SLAM approach is that it provides rich uncertainty representations that can include ambiguity due to data association. The belief computed using this method can be used to inform approaches in non-Gaussian active SLAM (as explored, for example using particle filter-based methods in [7]), or more generally for non-Gaussian belief space planning [45].

The latter proposed approach enables semantic SLAM with non-traditional sensing modalities. By choosing a representation that does not make assumptions about the measurement distribution, we are able to deal with ambiguous data associations

that arise from non-Gaussian sensor models, for example in the case of multiple returns by a sonar. This opens interesting new avenues for future work in object detection and tracking in underwater environments.

We assumed a simple geometric model and focused on comparison of data association methods. Another important area for future work is the application of our approach using novel geometric representations, e.g. quadrics [54, 41]. More detailed information about the geometric state of each landmark (e.g. the 6-degree-of-freedom pose) can be very helpful in disambiguating landmarks.

Finally, there is ample room for future work on the development of non-Gaussian inference methods for SLAM. The multimodal iSAM framework enables navigation with a much broader class of sensor models than typically accommodated by SLAM solutions. That said, nonlinear least-squares optimization approaches to SLAM have the benefit of nearly two decades of research and development in contrast to the multimodal iSAM approach. Consequently, new approaches to non-Gaussian inference for SLAM as well as optimizations of the current approaches are needed in order for non-Gaussian SLAM methods to reach the real-time performance and computational efficiency of competitive methods making nonlinear Gaussian measurement assumptions.

## 6.5 Concluding Remarks

The two methods presented in this thesis represent steps toward coupling data association and SLAM into a single navigation framework. Furthermore, we describe ways in which semantics can inform data association and be incorporated into the SLAM problem, while accommodating the discrete measurement noise due to detector misclassification. There has been substantial recent interest in three ideas in the context of semantic SLAM: "semantics informing SLAM," i.e. the use of semantics to aid in navigation, "SLAM informing semantics," leveraging knowledge of the geometric structure of the environment to perform, for example, better object detection, and "joint SLAM and semantics inference," which is the use of semantics and SLAM

to the mutual benefit of one another [6]. The methods presented in this thesis are important steps toward addressing the latter problem of joint SLAM and semantics inference, and consequently toward a unifying perspective on navigation and scene understanding.

# Appendix A

# Matrix Manifolds in SLAM

Robot poses $\boldsymbol{x}$ consist of a position component $\mathbf{t}$ and a rotation component $\mathbf{R}$. In two dimensions, the position of the vehicle is a real-valued vector in $\mathbb{R}^2$, and similarly in three-dimensions the position vector is in $\mathbb{R}^3$. These represent the translation of the robot with respect to some global coordinate system. Rotations in two- and three-dimensional coordinate systems, in contrast, have several possible representations, including quaternions $\mathbb{H}$, Euler angles (often denoted $\alpha, \beta, \gamma$ or "roll", "pitch", and "yaw"), or elements of the special orthogonal matrix groups SO(2) and SO(3) (for two and three dimensions, respectively). In this thesis, we consider rotations as elements of the matrix groups SO(2) and SO(3), and consequently robot poses $\boldsymbol{x}$ in the special Euclidean groups SE(2) and SE(3). Here we review several definitions and facts about these groups used implicitly in this thesis.

## A.1 Matrix Lie Groups and Lie Algebras

All of the groups we discuss are *Lie groups*, meaning they are also smooth, differentiable manifolds. A Lie group locally resembles a Euclidean vector space. In particular, all matrix Lie groups have a corresponding "infinitesimal group", termed a *Lie algebra*. The Lie algebra is a group in that it consists of a set and a binary operation, namely a vector space called the *tangent space* and a binary operation called the *Lie bracket* $[X, Y] = XY - YX$.

### A.1.1 Special Orthogonal Group

Representing two-dimensional rotations, the **special orthogonal group** in two dimensions SO(2) is defined as:

$$\mathrm{SO}(2) \triangleq \left\{ \mathbf{R} \in \mathbb{R}^{2\times 2} \mid \mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = I_{2\times 2},\ \det \mathbf{R} = 1 \right\}. \tag{A.1}$$

That is, it consists of $2 \times 2$ orthogonal matrices with determinant 1. SO(2) is isomorphic to the unit circle $\mathrm{S}^1$. This isomorphism is intuitive, since rotations in two-dimensional space can be equivalently expressed with a single yaw angle $\gamma$:

$$\mathrm{SO}(2) \triangleq \left\{ \begin{bmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{bmatrix} \;\middle|\; \gamma \in \mathbb{R}/2\pi\mathbb{Z} \right\}. \tag{A.2}$$

The angle $\gamma$ is real-valued (modulo integer multiples of $2\pi$).

The corresponding Lie algebra is $\mathfrak{so}(2)$:

$$\mathfrak{so}(2) \triangleq \left\{ \begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix} \;\middle|\; \theta \in \mathbb{R} \right\}. \tag{A.3}$$

Thus $\mathfrak{so}(2)$ consists of the vector space $\mathbb{R}$ and has a trivial Lie bracket $[X, Y] = XY - YX = 0$ since the matrices commute.

The special orthogonal group in three dimensions SO(3) is defined as:

$$\mathrm{SO}(3) \triangleq \left\{ \mathbf{R} \in \mathbb{R}^{3\times 3} \mid \mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = I_{3\times 3},\ \det \mathbf{R} = 1 \right\}. \tag{A.4}$$

It consists of $3 \times 3$ orthogonal matrices with determinant 1. In general, the $d$-dimensional special orthogonal group $\mathrm{SO}(d)$ group consists of the $d \times d$ orthogonal matrices with determinant 1.

The Lie algebra corresponding to the group SO(3) is denoted $\mathfrak{so}(3)$ and defined

as:

$$\mathfrak{so}(3) \triangleq \left\{ \begin{bmatrix} 0 & -\phi_3 & \phi_2 \\ \phi_3 & 0 & -\phi_1 \\ -\phi_2 & \phi_1 & 0 \end{bmatrix} \;\middle|\; \boldsymbol{\phi} = [\phi_1 \; \phi_2 \; \phi_3]^T \in \mathbb{R}^3 \right\}, \tag{A.5}$$

which is simply the set of skew-symmetric matrices in $\mathbb{R}^{3\times 3}$.

### A.1.2 Special Euclidean Group

The **special Euclidean group** in two dimensions SE(2) is defined as:

$$\mathrm{SE}(2) \triangleq \left\{ \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_2^T & 1 \end{bmatrix} \;\middle|\; \mathbf{R} \in \mathrm{SO}(2),\ \mathbf{t} \in \mathbb{R}^2 \right\}. \tag{A.6}$$

Thus, it consists of a rotation $\mathbf{R} \in \mathrm{SO}(2)$ and a translation $\mathbf{t} \in \mathbb{R}^2$. Here we denote the $n \times 1$ zero vector as $\mathbf{0}_n$.

The corresponding Lie algebra is $\mathfrak{se}(2)$:

$$\mathfrak{se}(2) \triangleq \left\{ \begin{bmatrix} 0 & -\theta & \rho_1 \\ \theta & 0 & \rho_2 \\ 0 & 0 & 0 \end{bmatrix} \;\middle|\; [\rho_1 \; \rho_2 \; \theta] \in \mathbb{R}^3 \right\} \tag{A.7}$$

In three dimensions, the special Euclidean group SE(3) is defined as:

$$\mathrm{SE}(3) \triangleq \left\{ \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_3^T & 1 \end{bmatrix} \;\middle|\; \mathbf{R} \in \mathrm{SO}(3),\ \mathbf{t} \in \mathbb{R}^3 \right\}. \tag{A.8}$$

The corresponding Lie algebra is $\mathfrak{se}(3)$:

$$\mathfrak{se}(3) \triangleq \left\{ \begin{bmatrix} \boldsymbol{\phi}^\wedge & \boldsymbol{\rho} \\ \mathbf{0}_3^T & 0 \end{bmatrix} \;\middle|\; \boldsymbol{\rho}, \boldsymbol{\phi} \in \mathbb{R}^3 \right\}, \tag{A.9}$$

where we use the "hat" notation $(\cdot)^\wedge$ to denote the $n \times n$ skew-symmetric matrix

consisting of elements of the $n$-dimensional vector argument. For example, for a 3-dimensional vector $\boldsymbol{a}$:

$$\boldsymbol{a}^{\wedge} \triangleq \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}. \tag{A.10}$$

We similarly define the inverse "vee" operation $(\cdot)^{\vee}$ for a $3 \times 3$ skew-symmetric matrix as:

$$\begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}^{\vee} \triangleq \boldsymbol{a}. \tag{A.11}$$

### A.1.3 Exponential and Logarithm Maps

The *exponential map* and *logarithm map* allow us to relate elements of a matrix Lie group to their correspondents in the Lie algebra. In particular, the exponential map produces an element of a matrix Lie group $\mathbf{M}$ from an element of the corresponding Lie algebra $\mathbf{A} = \boldsymbol{a}^{\wedge}$:

$$\mathbf{M} = \exp(\mathbf{A}) = \sum_{n=0}^{\infty} \frac{1}{n!} \mathbf{A}^n. \tag{A.12}$$

The logarithm map takes as argument an element $\mathbf{M}$ from the matrix Lie group and produces in turn the corresponding element $\mathbf{A}$ of the Lie algebra:

$$\mathbf{A} = \log(\mathbf{M}) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} (\mathbf{M} - I)^n, \tag{A.13}$$

where $I$ is the identity matrix with dimensionality equal to that of $\mathbf{M}$.

The exponential and logarithm maps are defined over matrices, but in the case of the exponential map, the "input" matrices we are concerned with are skew-symmetric, and for the logarithm map, the "output" matrices are skew-symmetric. For conve-

nience, we define the following notation for the exponential and logarithm maps:

$$\mathrm{Exp}(\boldsymbol{a}) \triangleq \exp(\boldsymbol{a}^{\wedge}) \tag{A.14}$$

$$\mathrm{Log}(\mathbf{A}) \triangleq \log(\mathbf{A})^{\vee}, \tag{A.15}$$

where here Exp is a function of an $n$-dimensional vector $\boldsymbol{a}$ and recovers a matrix, while Log is a function which operates on $n \times n$ matrices $\mathbf{A}$ and "returns" an $n$-dimensional vector.

## A.2 Operations on Poses

We make use of *pose composition* $\oplus$ and *pose inversion* $\ominus$ in this thesis. Here we briefly describe these operations as defined in [51], [4].

### A.2.1 Pose Composition

The composition of two poses $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathrm{SE}(d)$, denoted by $\oplus$ is defined as:

$$\boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \triangleq \begin{bmatrix} \mathbf{R}_1\mathbf{R}_2 & \mathbf{R}_1\mathbf{t}_2 + \mathbf{t}_1 \\ \mathbf{0}_d^T & 1 \end{bmatrix} \tag{A.16}$$

where $\mathbf{R}_1 \in \mathrm{SO}(d)$ and $\mathbf{t}_1 \in \mathbb{R}^d$ are the rotation matrix and translation vector for pose $\boldsymbol{x}_1$, respectively, and likewise for $\boldsymbol{x}_2$.

### A.2.2 Pose Inversion

Pose inversion, denoted $\ominus$ is defined as:

$$\ominus\boldsymbol{x} \triangleq \boldsymbol{x}^{-1} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_d^T & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{R}^T & -\mathbf{R}^T\mathbf{t} \\ \mathbf{0}_d^T & 1 \end{bmatrix}, \tag{A.17}$$

where $\mathbf{R} \in \mathrm{SO}(d)$ and $\mathbf{t} \in \mathbb{R}^d$ are the rotation matrix and translation vector for the pose $\boldsymbol{x}$.

# Bibliography


[1] BlueRobotics BlueROV2 Underwater Vehicle. `https://bluerobotics.com/store/rov/bluerov2/bluerov2/`. Accessed August 2019.

[2] Nikolay Atanasov, Menglong Zhu, Kostas Daniilidis, and George J Pappas. Semantic localization via the matrix permanent. In *Robotics: Science and Systems*, volume 2, 2014.

[3] Yaakov Bar-Shalom and Edison Tse. Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11(5):451–460, 1975.

[4] Timothy D Barfoot. *State Estimation for Robotics*. Cambridge University Press, 2017.

[5] Sean L Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J Pappas. Probabilistic data association for semantic SLAM. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1722–1729. IEEE, 2017.

[6] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[7] Luca Carlone, Jingjing Du, Miguel Kaouk Ng, Basilio Bona, and Marina Indri. Active slam and exploration with particle filters using kullback-leibler divergence. *Journal of Intelligent & Robotic Systems*, 75(2):291–311, 2014.

[8] Ingemar J Cox and John J Leonard. Probabilistic data association for dynamic world modeling: A multiple hypothesis approach. In *Advanced Robotics, 1991.'Robots in Unstructured Environments', 91 ICAR., Fifth International Conference on*, pages 1287–1294. IEEE, 1991.

[9] Ingemar J Cox and John J Leonard. Modeling a dynamic environment using a bayesian multiple hypothesis approach. *Artificial Intelligence*, 66(2):311–344, 1994.

[10] Frank Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012.

[11] Frank Dellaert, Michael Kaess, et al. Factor graphs for robot perception. *Foundations and Trends® in Robotics*, 6(1-2):1–139, 2017.

[12] Kevin Doherty, Genevieve Flaspohler, Nicholas Roy, and Yogesh Girdhar. Approximate distributed spatiotemporal topic models for multi-robot terrain characterization. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3730–3737. IEEE, 2018.

[13] Kevin Doherty, Dehann Fourie, and John J Leonard. Multimodal semantic SLAM with probabilistic data association. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[14] Hugh Durrant-Whyte, Somajyoti Majumder, Sebastian Thrun, Marc De Battista, and Steve Scheding. A Bayesian algorithm for simultaneous localisation and map building. In *Robotics Research*, pages 49–60. Springer, 2003.

[15] Arne D Ekstrom, Hugo J Spiers, Véronique D Bohbot, and R Shayna Rosenbaum. *Human spatial navigation*. Princeton University Press, 2018.

[16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[18] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics*, 33(1):1–21, 2016.

[19] Dehann Fourie. *Multi-modal and inertial sensor solutions for navigation-type factor graphs*. PhD thesis, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution, 2017.

[20] Dehann Fourie, John Leonard, and Michael Kaess. A nonparametric belief solution to the Bayes tree. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 2189–2196. IEEE, 2016.

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[23] Andreas Geiger, Julius Ziegler, and Christoph Stiller. StereoScan: Dense 3d Reconstruction in Real-time. In *IEEE Intelligent Vehicles Symposium*, Baden-Baden, Germany, June 2011.

[24] Michael Grupp. evo: Python package for the evaluation of odometry and slam. `https://github.com/MichaelGrupp/evo`, 2017.

[25] Dirk Hähnel, Sebastian Thrun, Ben Wegbreit, and Wolfram Burgard. Towards lazy data association in slam. In *Robotics Research. The Eleventh International Symposium*, pages 421–431. Springer, 2005.

[26] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[27] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7310–7311, 2017.

[28] Michael Kaess and Frank Dellaert. Covariance recovery from a square root information matrix for data association. *Robotics and autonomous systems*, 57(12):1198–1210, 2009.

[29] Michael Kaess, Viorela Ila, Richard Roberts, and Frank Dellaert. The bayes tree: An algorithmic foundation for probabilistic robot mapping. In *Algorithmic Foundations of Robotics IX*, pages 157–173. Springer, 2010.

[30] Michael Kaess, Hordur Johannsson, Richard Roberts, Viorela Ila, John J Leonard, and Frank Dellaert. iSAM2: Incremental smoothing and mapping using the Bayes tree. *The International Journal of Robotics Research*, 31(2):216–235, 2012.

[31] Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[32] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.

[33] Yasir Latif, César Cadena, and José Neira. Robust loop closing over time for pose graph slam. *The International Journal of Robotics Research*, 32(14):1611–1626, 2013.

[34] John J Leonard and Hugh F Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on robotics and Automation*, 7(3):376–382, 1991.

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[36] John McCormac, Ronald Clark, Michael Bloesch, Andrew J Davison, and Stefan Leutenegger. Fusion++: Volumetric Object-Level SLAM. *arXiv preprint arXiv:1808.08378*, 2018.

[37] Michael Montemerlo, Sebastian Thrun, Daphne Koller, and Ben Wegbreit. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proc. of the AAAI National Conference on Artificial Intelligence, 2002*, 2002.

[38] Beipeng Mu, Shih-Yuan Liu, Liam Paull, John Leonard, and Jonathan P How. SLAM with objects using a nonparametric pose graph. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 4602–4609. IEEE, 2016.

[39] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[40] José Neira and Juan D Tardós. Data association in stochastic mapping using the joint compatibility test. *IEEE Transactions on robotics and automation*, 17(6):890–897, 2001.

[41] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. QuadricSLAM: Constrained Dual Quadrics from Object Detections as Landmarks in Semantic SLAM. *IEEE Robotics and Automation Letters (RA-L)*, 2018.

[42] Edwin Olson and Pratik Agarwal. Inference on networks of mixtures for robust robot mapping. *The International Journal of Robotics Research*, 32(7):826–840, 2013.

[43] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[44] Max Pfingsthorn and Andreas Birk. Simultaneous localization and mapping with multimodal probability distributions. *The International Journal of Robotics Research*, 32(2):143–171, 2013.

[45] Robert Platt, Leslie Kaelbling, Tomas Lozano-Perez, and Russ Tedrake. Non-gaussian belief space planning: Correctness and complexity. In *2012 IEEE International Conference on Robotics and Automation*, pages 4711–4717. IEEE, 2012.

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[47] Donald Reid. An algorithm for tracking multiple targets. *IEEE transactions on Automatic Control*, 24(6):843–854, 1979.

[48] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013.

[49] Aleksandr V Segal and Ian D Reid. Hybrid inference optimization for robust pose graph estimation. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2675–2682. IEEE, 2014.

[50] Christopher M Smith and John J Leonard. A multiple-hypothesis approach to concurrent mapping and localization for autonomous underwater vehicles. In *Field and Service Robotics*, pages 237–244. Springer, 1998.

[51] Randall Smith, Matthew Self, and Peter Cheeseman. Estimating uncertain spatial relationships in robotics. In *Autonomous robot vehicles*, pages 167–193. Springer, 1990.

[52] Erik B Sudderth, Alexander T Ihler, Michael Isard, William T Freeman, and Alan S Willsky. Nonparametric belief propagation. *Communications of the ACM*, 53(10):95–103, 2010.

[53] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018.

[54] Niko Sünderhauf and Michael Milford. Dual quadrics from object detection bounding boxes as landmark representations in SLAM. *arXiv preprint arXiv:1708.00965*, 2017.

[55] Niko Sünderhauf and Peter Protzel. Switchable constraints for robust pose graph SLAM. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 1879–1884. IEEE, 2012.

[56] Jinkun Wang and Brendan Englot. Robust exploration with multiple hypothesis data association. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3537–3544. IEEE, 2018.

[57] Steven Xiaogang Wang. *Maximum weighted likelihood estimation*. PhD thesis, University of British Columbia, 2001.

[58] Shichao Yang and Sebastian Scherer. CubeSLAM: Monocular 3D object detection and SLAM without prior models. *arXiv preprint arXiv:1806.00557*, 2018.