

# Semantic Semi-Incremental Data-Association-Free Object SLAM

Yihao Zhang<sup>1\*</sup>, Jungseok Hong<sup>1\*</sup>, John J. Leonard<sup>1</sup>

**Abstract**—Data association between landmark measurements and landmark variables has long been a central challenge in SLAM, as estimation accuracy depends critically on associating measurements with the correct landmark variables. Recent advances in deep learning have created new opportunities for the problem; data association can now leverage not only positional measurements but also semantic information about object landmarks, such as class labels from neural object detectors and feature vectors from visual foundation models. In this paper, we present a generalized data-association-free SLAM framework that jointly estimates data associations, robot poses, landmark positions, and landmark semantics from odometry, and positional and semantic measurements of landmarks. The proposed framework (i) creates a synergy between data association and landmark semantics estimation; (ii) adopts a semi-incremental estimation scheme for improved accuracy and computational efficiency; and (iii) provides a principled justification, guidelines, and heuristics for landmark-number estimation, improving the interpretability and practical usability of the framework. The proposed framework and algorithms are evaluated on synthetic and real-world datasets with two types of semantic information, class labels and real-valued feature vectors, and demonstrate superior performance compared to strong baselines.

## I. INTRODUCTION

Object SLAM builds a map with objects as landmarks. It is a step beyond traditional geometry-only SLAM toward semantic scene understanding. Objects are typically detected by a neural module, which also provides semantic information about them. Recent work [1]–[4] explores the use of such semantic information to facilitate data association of landmark measurements. Our work advances this line of research. In particular, building on the data-association-free landmark-based SLAM (DAF-SLAM) framework, we develop a semantic data-association-free object SLAM framework that jointly estimates robot poses, landmark positions, landmark semantics, data associations, and the number of landmarks given odometry measurements, landmark position measurements, and landmark semantic measurements.

DAF-SLAM [5] studies a minimal problem in which semantic information is not available; therefore, data association has to be inferred solely from odometry and landmark position measurements. The authors recognize that when data association is unknown, the number of landmarks in the environment is also unknown and may be as large as the number of landmark measurements received. They split the problem into an inner problem that estimates the data associations given the number of landmarks and an outer

problem that estimates the number of landmarks, assuming the inner problem can be solved. Various algorithms have been developed to solve these sub-problems.

While DAF-SLAM establishes a promising foundation, extending it to an efficient and accurate object SLAM framework introduces several important challenges. First, the DAF-SLAM framework does not exploit semantic information that is often available from modern object detectors and visual foundation models. Second, its batch optimization strategy can suffer from accumulated odometry drift when pose estimates are initialized from odometry over long trajectories. Third, a deeper understanding of the landmark-number estimation mechanism is desirable for formulation interpretation and parameter selection.

In this paper, we extend the DAF-SLAM framework and address the aforementioned challenges through four main contributions. (i) We provide a principled analysis of the landmark-number estimation mechanism based on local optimality conditions. (ii) We develop a semi-incremental (or block-incremental) processing scheme inspired by the classical divide-and-conquer strategy to reduce drift and improve computational efficiency. (iii) We integrate semantic information into the estimation process, enabling joint inference of data associations and landmark semantics. The semantic measurements may take the form of either real-valued feature vectors or one-hot vectors, which are compatible with the outputs of visual foundation models and object detectors. (iv) We evaluate the proposed algorithms extensively on synthetic datasets, a real-world dataset using an object detector to extract semantic measurements, and a real-world underwater dataset using a visual foundation model to extract semantic measurements. The proposed algorithms outperform strong baselines across these diverse datasets. Together, these contributions advance the development of a more general and efficient data-association-free object SLAM framework.

## II. RELATED WORK

Research on data association in SLAM has a long history. Neira and Tardós [6] and Montemerlo et al. [7] provide an early review of data association methods in SLAM, including maximum likelihood, joint compatibility branch and bound, combined constraint data association, iterative closest point, and multi-hypothesis tracking. Traces of these methods can still be seen in modern geometry-only and object SLAM methods. For example, object SLAM pipelines [2], [3] belong to the maximum likelihood category. They compute the Mahalanobis distance, derived from maximum likelihood estimation, to handle the positional component, and the cosine similarity (or inner product) to handle

<sup>1</sup>Yihao Zhang, Jungseok Hong, and John Leonard are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology ({yihaozh, jungseok, jleonard}@mit.edu).  
\* indicates equal contribution.

the semantic feature-vector component of data association. Iterative-style methods such as Max-Mixture [1], DC-SAM [8], and DAF-SLAM [5] are, in principle, similar to iterative closest point or coordinate descent. However, how variables are split can affect the algorithm performance. Both Max-Mixture and DC-SAM split the variables into discrete (i.e., data associations) and continuous (i.e., landmark positions and robot poses) variables, and solve them separately in an alternating fashion. Whereas DAF-SLAM considers a smarter split, where one step solves data associations and landmark positions, and another step solves robot poses and again landmark positions. Mature algorithms exist to solve both sub-problems, yielding improved performance. A recent approach [9] solves the discrete-continuous problem with semi-definite relaxation but has limited scalability.

Although still running maximum likelihood or maximum a posteriori estimation, a distinct line of object SLAM work [10]–[12] models data association as a Dirichlet process so that landmarks with more associated measurements are more likely to be associated again. Another line of object SLAM work [4], [13] takes an expectation over data associations in an expectation-maximization style to eliminate the discrete data association variables in the inference.

When building object SLAM systems, researchers often employ sensor-dependent front-end association methods. For example, SLAM++ [14] uses image-based projective data association; room numbers on door signs are used for association in [15]; ODAM [16] (mapping-only) makes associations with a graph neural network that takes in different image-based attributes and features. These front-end methods can be easily incorporated as a module in a SLAM system. However, mistakes made in these front-end associations are not recoverable in the back end. In addition, a redesign of the data association module is required for these sensor-dependent methods when the sensor is changed. In this paper, we focus on estimating associations in the sensor-agnostic back-end solver. In particular, we extend and enhance the DAF-SLAM framework [5] to develop a more general and efficient object SLAM framework.

### III. BACKGROUND AND DAF-SLAM RECAP

The data-association-free SLAM (DAF-SLAM) problem is named in [5] to indicate a setup where only odometry and landmark measurements are available to a SLAM solver, but not data associations of the landmark measurements. Mathematically, we denote the robot trajectory as  $\mathbf{x} = \{x_i\}_{i=1}^N$ , where  $x_i \in \text{SE}(d)$  is the unknown  $i$ -th pose of the robot in dimension  $d = 2$  or  $3$ , and  $x_i = (R_i, t_i)$  are the rotation and translation components. The unknown landmark positions are denoted by  $\mathbf{y} = \{y_j\}_{j=1}^K$  where  $y_j \in \mathbb{R}^d$ . Finally, the given landmark measurements are  $\tilde{\mathbf{z}} = \{\tilde{z}_k\}_{k=1}^M$ .  $\tilde{z}_k \in \mathbb{R}^d$  is assumed to be generated from the standard model:

$$\tilde{z}_k = R_{i_k}^\top (y_{j_k} - t_{i_k}) + \epsilon_k \quad (1)$$

where a measurement  $\tilde{z}_k$  subject to zero-mean Gaussian noise  $\epsilon_k \sim \mathcal{N}(0, \Sigma)$  is made in the local frame of robot pose  $i_k$ . This  $k$ -th measurement can be associated with

the corresponding pose  $x_{i_k}$  via timestamp synchronization. However, the data association, i.e., the index of the landmark being measured ( $j_k$ ), is unknown. Furthermore, since data associations are unknown, the number of landmarks ( $K$ ) in the environment is also unknown.

It is proposed in [5] to split the DAF-SLAM problem into an inner problem and an outer problem. The inner problem takes the form:

$$\min_{\substack{\mathbf{x} \in \text{SE}(d)^N \\ \mathbf{y} \in \mathbb{R}^{d \times K}}} f_{\text{odom}}(\mathbf{x}) + \sum_{k=1}^M \min_{j_k \in [K]} \|R_{i_k}^\top (y_{j_k} - t_{i_k}) - \tilde{z}_k\|_\Sigma^2 \quad (2)$$

where  $[K] = \{1, \dots, K\}$ . (2) is a function of the number of landmarks ( $K$ ). Given  $K$ , it minimizes the residuals of landmark measurements over data associations ( $j_k$ ), which can be viewed as a max-mixture model [17]. (2) is solved by alternating minimization between solving  $\{j_k, \mathbf{y}\}$  through k-means and k-means++ [18] and solving  $\{\mathbf{x}, \mathbf{y}\}$  through a SLAM solver [19]. This alternating scheme is shown to be advantageous over alternating between  $j_k$  alone and  $\{\mathbf{x}, \mathbf{y}\}$ . In addition, k-means++ eliminates the need of an initial guess for  $\mathbf{y}$  so only odometry is needed to initialize  $\mathbf{x}$ .

Let the optimal objective value of (2) be  $f_{\text{slam}}^*(K)$ . The outer problem of estimating the number of landmarks ( $K$ ) is

$$\min_{K \in \{1, \dots, M\}} f_{\text{slam}}^*(K) + \beta K \quad (3)$$

where  $\beta$  is a hyperparameter. (3) is solved with multi-resolution gridding, which is a zeroth-order method that uses the inner problem (2) solver to evaluate  $f_{\text{slam}}^*(K)$ . The full algorithm is named kSLAM for its use of k-means.

## IV. METHOD

### A. Local Optimality Condition

One important but missing analysis in [5] is the local optimality condition for (3). This optimality condition not only plays a role in finding heuristics for setting  $\beta$  but also helps explain the working principle of the outer problem formulation. Suppose the actual number of landmarks is  $K^*$ . In order for the minimizer of (3) to be  $K^*$ , at least the following necessary local conditions have to be met:

$$f_{\text{slam}}^*(K^* + 1) + \beta \geq f_{\text{slam}}^*(K^*) \quad (4)$$

$$f_{\text{slam}}^*(K^* - 1) - \beta \geq f_{\text{slam}}^*(K^*) \quad (5)$$

Combining (4) and (5), we can derive a condition on  $\beta$ :

$$f_{\text{slam}}^*(K^* - 1) - f_{\text{slam}}^*(K^*) \geq \beta \geq f_{\text{slam}}^*(K^*) - f_{\text{slam}}^*(K^* + 1) \quad (6)$$

$f_{\text{slam}}^*(K^* - 1)$  is the optimal SLAM objective value with one landmark variable fewer than the actual number so the measurements for this missing landmark have to be associated incorrectly to other landmarks (if we assume that  $f_{\text{slam}}^*(K^*)$  contains all correct associations). The residuals  $R_{i_k}^\top (y_{j_k} - t_{i_k}) - \tilde{z}_k$  for these measurements do not follow the Gaussian model since they are caused by association errors rather than sensor noise. When the inter-landmark spacing is large relative to the landmark measurement noise (Fig. 1a),

mis-associating measurements with a distant wrong landmark can cause large residuals in the objective function. This is where formulation (3) works robustly (see the visualization of  $f_{\text{slam}}^*(K)$  in the Appendix). In this case, the large upper bound  $\beta_{\text{ub}} = f_{\text{slam}}^*(K^* - 1) - f_{\text{slam}}^*(K^*)$  creates a favorable gap above the lower bound  $\beta_{\text{lb}} = f_{\text{slam}}^*(K^*) - f_{\text{slam}}^*(K^* + 1)$  for  $\beta$  to satisfy (6). Since any mis-association is unlikely to be in the optimal solution, the one extra landmark in  $f_{\text{slam}}^*(K^* + 1)$  would only split the measurements of a landmark. Under additional assumptions (see details in the Appendix), such as moderate odometry noise and negligible residual changes from the other measurement terms, the reduction in the objective value (i.e.,  $\beta_{\text{lb}}$ ) due to the split is therefore capped by the maximum residuals of all measurements associated with a landmark:

$$\beta_{\text{lb}} \leq \max_{j_k^*} \sum_{k \in \mathcal{M}(j_k^*)} \|R_{i_k}^{*T} (y_{j_k^*}^* - t_{i_k}^*) - \bar{z}_k\|_{\Sigma}^2 \quad (7)$$

where  $\mathcal{M}(j_k^*)$  is the set of measurements associated with landmark  $j_k^*$  and all the optimal values are computed at  $f_{\text{slam}}^*(K^*)$ .

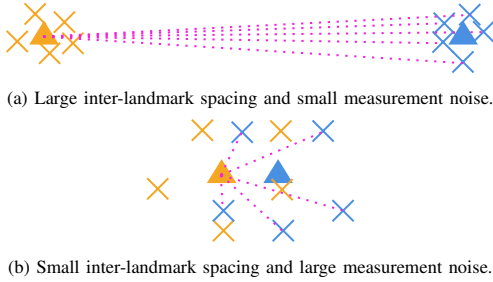


Fig. 1: Pre-optimization errors induced by mis-associations after removing the blue landmark. The crosses are measurements of landmarks. The length of each dashed magenta line is the magnitude of a mis-association error.

While the upper bound  $\beta_{\text{ub}}$  is high, to ensure that  $\beta$  is greater than  $\beta_{\text{lb}}$ , the right-hand side of (7) can be used as a heuristic to set  $\beta$  so as to satisfy (6). Although  $\arg\max_{j_k^*}$  and  $\mathcal{M}(j_k^*)$  are not known beforehand, we can estimate the number of measurements for a landmark ( $|\hat{\mathcal{M}}(j_k^*)|$ ) from prior information such as average robot speed, field of view, sensor range, and sensor frame rate (see the Appendix). Since  $\sum_{k \in \mathcal{M}(j_k^*)} \|R_{i_k}^{*T} (y_{j_k^*}^* - t_{i_k}^*) - \bar{z}_k\|_{\Sigma}^2$  follows a  $\chi^2$  distribution due to (1),  $\beta$  can be heuristically set by the inverse  $\chi^2$  distribution with  $d \times |\hat{\mathcal{M}}(j_k^*)|$  degrees of freedom and a sufficiently large probability threshold to be robust to violations of the assumptions underlying (7). The preceding analysis provides a more principled explanation of the heuristic in [5].

When the inter-landmark spacing is small relative to the landmark measurement noise, measurements from different landmarks may mix together (Fig. 1b). Mis-associating measurements with a nearby landmark incurs a cost similar to the original sensor noise error. Thus,  $f_{\text{slam}}^*(K^* - 1)$  is close to  $f_{\text{slam}}^*(K^*)$ , and the upper bound  $\beta_{\text{ub}}$  is lower. In the extreme case where two landmarks are on top of each other, removing one does not increase the objective value. Therefore,  $\beta$  is set empirically in this regime because setting

$\beta$  using the previous heuristic (7) may cause it to exceed  $\beta_{\text{ub}}$ . However, a starting value for  $\beta$  ( $\frac{2|\hat{\mathcal{M}}(j_k^*)|}{\pi}$ ) based on quantitative modeling of the effect of adding a landmark ( $f_{\text{slam}}^*(K^* + 1)$ ) by splitting Gaussian samples, along with additional guidelines provided in the Appendix, is used in an experiment (Section V-B) where landmarks are considered clustered relative to the measurement noise.

### B. Incorporating Semantic Feature Vectors

DAF-SLAM studies the setting where only spatial measurements of landmarks are available. However, for applications such as object SLAM, we may also have landmark feature vectors predicted by an object detector (e.g., one-hot classification vectors from Mask R-CNN [20]) or a foundation model (e.g., real-valued vectors from DINO [21]). To enable synergy between data association and feature vector estimation, we incorporate these vectors by assuming a Gaussian noise model (which may be relaxed in practice in Section IV-D):

$$\bar{z}_{s,k} = y_{s,j_k} + \epsilon_{s,k} \quad (8)$$

where  $\bar{z}_{s,k} \in \mathbb{R}^s$  is the semantic measurement (i.e., a feature vector of length  $s$ ) of the landmark semantics ( $y_{s,j_k} \in \mathbb{R}^s$ ), corrupted by zero-mean Gaussian noise  $\epsilon_{s,k} \sim \mathcal{N}(0, \Sigma_s)$ . The inner problem (2) is modified to include a semantic residual:

$$\min_{\substack{\mathbf{x} \in \text{SE}(d)^N \\ \mathbf{y} \in \mathbb{R}^{d \times K} \\ \mathbf{y}_s \in \mathbb{R}^{s \times K}}} f_{\text{odom}}(\mathbf{x}) + \sum_{k=1}^M \min_{j_k \in [K]} \left\{ \|R_{i_k}^T (y_{j_k} - t_{i_k}) - \bar{z}_k\|_{\Sigma}^2 + w_s^2 \|y_{s,j_k} - \bar{z}_{s,k}\|_{\Sigma_s}^2 \right\} \quad (9)$$

where  $w_s^2$  is a weighting term used to normalize the feature-vector residual by length (if the residual scales with length) and to have a tunable balance between the spatial residual and the semantic residual. In the alternating minimization, the association step now solves the new objective function:

$$\min_{\substack{\mathbf{y} \in \mathbb{R}^{d \times K} \\ \mathbf{y}_s \in \mathbb{R}^{s \times K}}} \sum_{k=1}^M \min_{j_k \in [K]} \left\| \begin{bmatrix} R_{i_k} \bar{z}_k + t_{i_k} \\ \frac{\sigma}{\sigma_s} w_s \bar{z}_{s,k} \end{bmatrix} - \begin{bmatrix} y_{j_k} \\ \frac{\sigma}{\sigma_s} w_s y_{s,j_k} \end{bmatrix} \right\|^2 \quad (10)$$

where we assume isotropic noise covariances ( $\Sigma = \text{diag}(\sigma)$  and  $\Sigma_s = \text{diag}(\sigma_s)$ ) to simplify the exposition, although the anisotropic case can be derived in a similar fashion. In practice,  $\sigma_s$  is often unknown so  $\frac{\sigma}{\sigma_s} w_s$  can be lumped into a single hyperparameter. Equation (10) stacks the spatial and semantic residuals so that they can be solved jointly in a single pass of k-means (with k-means++ [18]). The data association ( $j_k$ ) becomes the cluster assignment, and the cluster centers are the landmark positions ( $y_{j_k}$ ) concatenated with the weighted feature vectors ( $\frac{\sigma}{\sigma_s} w_s y_{s,j_k}$ ).

In the estimation step, we take the data association ( $j_k$ ) results from the association step and solve  $\{\mathbf{x}, \mathbf{y}\}$  using a SLAM solver [19]. For the semantic part, once the data associations are given, it is decoupled from  $\{\mathbf{x}, \mathbf{y}\}$  in (9) so it can be solved independently. The solution to

$\min_{y_s} \sum_{k=1}^M \|y_{s,j_k} - \bar{z}_{s,k}\|^2$  is simply the average of  $\bar{z}_{s,k}$ 's associated with each  $y_{s,j_k}$ , which is already computed as the cluster center in the k-means step.

For the outer problem, we retain the original objective function (3) because  $\beta$  is set according to the Gaussian assumption of the generative process (Section IV-A), and the Gaussian assumption for the semantic part (8) may be less reliable than that for the positional part (1). Therefore, the semantic part does not enter the outer problem.

**Remark.** The feature vector estimation can be a simple add-on to the original algorithm. Instead of performing k-means in the  $d$ -dimensional space, we now run it in the  $(d+s)$ -dimensional space. This allows both spatial information and semantic information to facilitate data association during the association step, while the association results can in turn help better estimate the feature vectors of landmarks. For example, a semantically erroneous feature vector may be corrected if the measurement is correctly associated based on its positional component. The updated algorithm is named skSLAM for its additional semantic component.

### C. Semi-Incremental Processing

The original DAF-SLAM work [5] adopts a batch processing scheme so that the entire trajectory is processed all at once and all the data associations are jointly determined, thereby avoiding the difficulty of correcting wrong past data associations in a pure incremental scheme. However, drift accumulates along a trajectory and leads to large errors in the initial guess for  $\mathbf{x}$  (i.e., the odometry chain) toward the end of the trajectory when the trajectory is long or the odometry noise is high, causing lower accuracy for the batch scheme.

To reduce drift, we propose a semi-incremental scheme following the divide-and-conquer strategy. Specifically, we define intervals of robot poses indexed by  $n$ ,  $\mathbf{x}_n = \{x_i\}_{i=(n-1)N_n+1}^{\min(nN_n, N)}$ , where  $n \in \{1, \dots, \lceil \frac{N}{N_n} \rceil\}$  and  $N_n \in \{1, \dots, N\}$  is the interval length. These intervals divide the trajectory into  $\lceil \frac{N}{N_n} \rceil$  segments. We first run skSLAM on each segment independently (either sequentially if the measurements arrive sequentially or in parallel if they are available at once). In the heuristic for  $\beta$  (Section IV-A), the number of measurements is taken as  $N_n$  for the segment processing if  $|\hat{\mathcal{M}}(j_k^*)| > N_n$ , because  $N_n$  poses can at most produce  $N_n$  observations of a landmark. Once all segments have been processed, we chain them together by connecting the last pose of a segment to the first pose of the next segment using odometry and propagating all pose updates along the chain. We then perform global optimization with skSLAM on the chained segments, which now represent the entire trajectory (Fig. 2). The key saving is that the outer problem solution space over the entire trajectory is reduced from  $K \in \{1, \dots, M\}$  to  $\{1, \dots, \sum_n \hat{K}_n\}$ , where  $\hat{K}_n$  is the estimated number of landmarks for segment  $n$ . We name the updated algorithm skid-SLAM with “i” standing for (semi-)incremental and “d” standing for divide-and-conquer. It is semi-incremental because poses are processed block-wise in segments.

**Remark.** The segment processing first reduces drift within each segment. Therefore, the initial guess for the trajectory in the global optimization contains less accumulated error than the raw odometry chain. This scheme is analogous to the local and global bundle-adjustment in ORB-SLAM [22]. In terms of computation cost, the outer problem solver originally requires  $\mathcal{O}(\log(M))$  evaluations of  $f_{\text{slam}}^*(K)$ , each of which requires solving a nonlinear SLAM problem multiple times in the alternating minimization steps. The SLAM solver complexity can be as high as cubic in the problem size. Overall, this is computationally expensive and becomes especially worse as the number of poses ( $N$ ) and the number of measurements ( $M$ ) increase. By performing the proposed semi-incremental scheme, we have  $\lceil \frac{N}{N_n} \rceil$  small problems that keep the  $\mathcal{O}(\log(M_n))$  and cubic scaling low. Additionally, the global optimization can take advantage of the already estimated numbers of landmarks for the segments  $\sum_n \hat{K}_n$ , which is typically much smaller than  $M$  because each landmark usually has tens or hundreds of measurements for a high-frame-rate sensor such as a camera. The global optimization does not inherit the association results from the segment processing to avoid convergence to local minima.

### D. Practical Considerations

**One-hot feature vectors.** In practice, researchers may want to use one-hot classification vectors from an object detector (e.g., Mask R-CNN [20]) to describe object landmarks. Although our formulation and algorithm are derived from the real-valued feature vector assumption, our algorithm also works functionally for one-hot vectors. During the k-means association step (Section IV-B), the use of one-hot vectors would violate the Gaussian residual assumption (8) and the physical meaning of  $\sigma_s$ , but  $\frac{\sigma}{\sigma_s} w_s$  can be treated as a hyperparameter to balance the spatial and semantic terms. In the estimation step, taking the average of the one-hot vectors associated with a landmark and selecting the argmax element in the average vector is equivalent to majority voting for the landmark class. Therefore, the same algorithm can be applied to one-hot feature vectors without any modification.

**Parallel global optimization.** The global optimization (Section IV-C) can run in parallel with the segment processing. This parallel scheme runs the global optimization on a different thread whenever there are more than two processed segments available. The output of the parallel global optimization is stored as a segment so that it will be combined with the subsequent segments and jointly optimized together with them in the next global optimization. Fig. 2 shows a schematic of the different processing schemes. The parallel scheme is advantageous if the up-to-date pose estimates, incorporating the latest global optimization results, are required in real time. In principle, it should also be more accurate than the series scheme (Section IV-C) because drift is reduced further through the parallel global optimization before the final whole-trajectory global optimization. This parallel implementation is named skill-SLAM, with “ll” standing for parallel.

**Detector mistakes.** Object detectors, and similarly visual

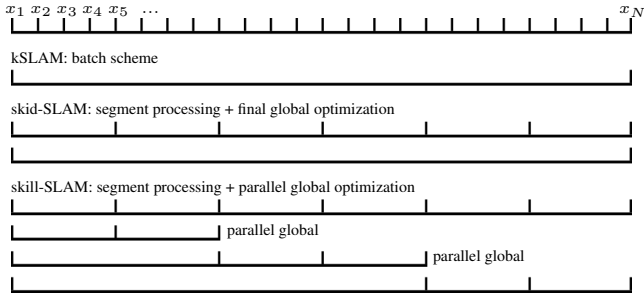


Fig. 2: Different processing schemes.

foundation models, may make three types of errors: detecting an irrelevant object as one of the target object classes (false positives), missing detections, and mislabeling objects as one of the other target classes (wrong labels). We discuss their effects and possible treatments. Sporadic false positives that are spatially close to true positives tend to be merged with the true positives (Section V-C) because the reduction in  $f_{\text{slam}}^*(K)$  by forming a new landmark for them usually does not exceed the penalty  $\beta$  in (3). Handling false positives is an interesting direction for future work. For occasional missed detections, if their corresponding landmark can still be estimated from detections at other viewing angles, we can render the estimated object in the frames where it is missed to correct the missing detections in post-processing. For mislabeling, a similar assumption is that the object is correctly labeled in the majority of frames from other viewing angles and correctly associated using the spatial information. In such a case, our algorithm can natively correct incorrect labels by selecting the most-voted class as the final semantic class estimate.

## V. EXPERIMENTS

### A. Synthetic Datasets and Ablation Study

Benchmarking and ablation experiments are conducted on the same 2D and 3D synthetic grid datasets as those in [5], but with the addition of one-hot and real-valued semantic feature vectors. We compare against three baselines, (i) raw odometry (ii) an alternating minimization algorithm (named Oracle) that is given the ground-truth number of landmarks and an initial guess for each landmark [5], [8], and (iii) the original batch processing kSLAM algorithm [5]. We also evaluate two ablated versions of our proposed algorithms, (i) running kSLAM repetitively every  $N_n$  poses on all poses seen so far to (semi-)incrementally process the trajectory (ikSLAM), and (ii) enabling the semantic feature vector processing (Section IV-B) on top of ikSLAM (iskSLAM). The expectation is that ikSLAM should achieve higher accuracy than kSLAM because odometry drift is reduced by incrementally rerunning the algorithm instead of performing one-time batch processing. iskSLAM may achieve the highest accuracy among all since it is the most redundant version. However, both ikSLAM and iskSLAM are computationally expensive and even prohibitive on large datasets due to the naive incremental rerunning scheme. We increase the number of inner solver iterations for all kSLAM

variants on the synthetic datasets to improve convergence and obtain a more accurate characterization of the trend. The performance of all methods is measured in terms of absolute trajectory error (ATE) after SIM(3) alignment between the estimated trajectory and a reference trajectory, which is obtained by optimizing the same measurements given the ground-truth data associations. We typically find that ATE is positively correlated with other metrics such as landmark position error. We sweep various dataset parameters in Fig. 3 – 6. Each data point in the plots is the median result of 20 simulations (different realizations of the noise) and the shaded region shows the 25% – 75% quartile range. We use the heuristic in Section IV-A to set  $\beta$ . Detailed dataset settings, algorithm parameter settings, and evaluation protocols are in the Appendix.

**Odometry noise.** The proposed semi-incremental scheme (Section IV-C) is designed to reduce drift while keeping the computational load low. In Fig. 3, we show that all the semi-incremental methods (ikSLAM, iskSLAM, skid-SLAM, and skill-SLAM) generally outperform the batch kSLAM algorithm. Further, the incorporation of semantic feature vectors (iskSLAM, skid-SLAM, and skill-SLAM) results in even lower ATE, particularly when the odometry noise is high. The baseline alternating minimization algorithm (Oracle) is almost always inferior, which is consistent with the results reported in [5].

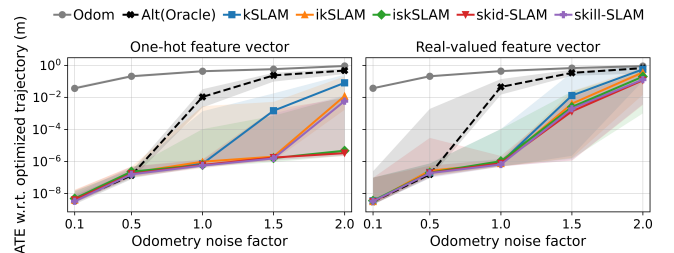


Fig. 3: Absolute trajectory error (ATE) in log scale for different odometry noise levels on the 3D grid dataset. The noise factor is multiplied to the default noise level (translational std = 0.05 (m) and rotational std = 0.005).

**Runtime.** Although the naive incremental rerunning scheme (ikSLAM and iskSLAM) also reduces ATE, it has worse runtime (Fig. 4). Among the algorithms that incorporate semantic information (iskSLAM, skid-SLAM, and skill-SLAM), iskSLAM is clearly the slowest. skill-SLAM is slower than skid-SLAM because the parallel global-optimization thread has to wait for the last global optimization to finish before starting the final whole-trajectory optimization, whereas skid-SLAM starts the whole-trajectory optimization immediately after the segment processing is completed. Among all the algorithms, Oracle is the fastest because it is given the ground-truth number of landmarks and does not spend time estimating it. On the smaller 3D grid dataset (216 poses), kSLAM is faster than our proposed algorithms. However, on the larger 2D grid dataset (500 poses), the savings from our divide-and-conquer strategy (Section IV-C) become apparent and enable the proposed algorithms to match or surpass the other baselines, particularly when

more landmarks are present. The runtime advantage becomes even more significant when the number of measurements per landmark is higher ( $|\mathcal{M}(j_k^*)| = 10$  here), resulting in a large total number of measurements ( $M$ ) (Section V-B).

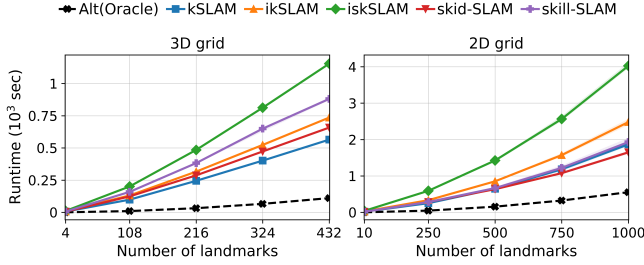


Fig. 4: Runtime for different numbers of landmarks.

**Semantic noise.** We vary the nominal detector mislabeling probability for one-hot semantic vectors (Fig. 5) and the Gaussian noise standard deviation for real-valued semantic feature vectors (Fig. 6). The nominal detector mislabeling probability specifies the probability with which each semantic measurement is intentionally corrupted. The “Raw” curve in each figure denotes the actual semantic error in the measurements after the noise is realized. The remaining curves show the errors of the estimated landmark semantics. The algorithms that process semantic information (iskSLAM, skid-SLAM, and skill-SLAM) generally achieve lower ATE than those do not (kSLAM and ikSLAM), but the benefits diminish as the semantic information becomes less reliable. On the other hand, joint estimation of semantics yields substantial reductions in landmark semantic estimation errors in both the one-hot and real-valued cases, as evidenced by the large differences between the raw semantic errors and the estimation errors.

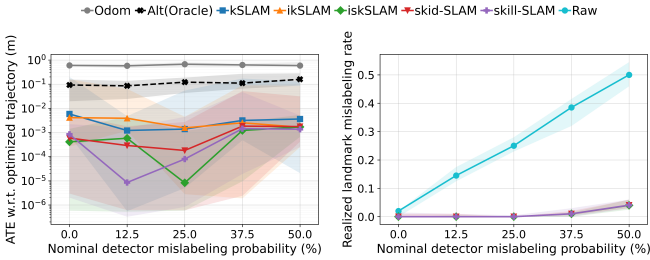


Fig. 5: Absolute trajectory error (ATE) in log scale and mislabeling rate for different nominal detector mislabeling probabilities on the 2D grid dataset with one-hot feature vectors.

### B. Kitchen Datasets with One-Hot Feature Vectors

We evaluate our algorithms on in-house indoor kitchen datasets and designed to test their ability to incorporate object detections and one-hot semantic vectors. We place various objects in a kitchen environment and remotely operate an RB-Y1 humanoid robot equipped with a ZED stereo camera, guiding it through the kitchen over multiple passes to mimic a household robot performing daily tasks. Two baselines are included: (i) the batch processing kSLAM algorithm

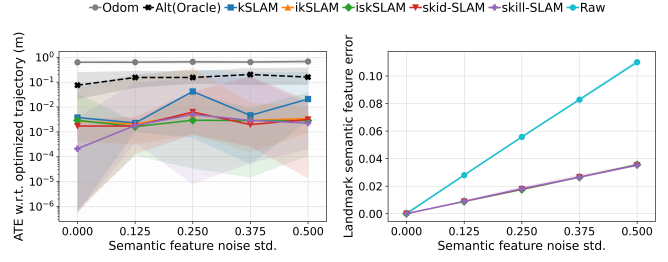


Fig. 6: Absolute trajectory error (ATE) in log scale and feature vector estimation error (Euclidean norm) for different semantic noise standard deviations on the 2D grid dataset with real-valued feature vectors.

[5], and (ii) an object-centric SLAM method [2]. The input odometry is computed using RGB-D visual odometry [23], [24]. The landmark measurements are taken as the centroids of segmented objects in point clouds back-projected from depth images, where the object masks are obtained from Mask R-CNN [20] for kSLAM and our proposed algorithms, and from YOLOv8 for the object-centric SLAM baseline [2]. To provide ground truth for evaluation, we place a unique AprilTag [25] close to each object (rather than on the object itself, to avoid interfering with the object detector). These tags serve as pseudo-landmarks whose measurements are extracted in a similar manner from the back-projected depth images. The identities of the tags allow us to associate the tag measurements and run standard landmark-based SLAM. The resulting estimated trajectory is used as the pseudo-ground-truth trajectory for evaluation, and the estimated tag positions serve as qualitative references for the object positions because they are slightly offset from the objects, making them unsuitable for quantitative evaluation).

The objects are strategically arranged in two layouts: (i) clusters of objects from different classes, and (ii) clusters of objects from the same class. As discussed in Section IV-A, for environments of this type, where landmarks are close to one another relative to the landmark measurement noise,  $\beta$  is set empirically. However, following the modeling results and guidelines presented in the Appendix, a reasonable starting value for  $\beta$  is  $\frac{2|\mathcal{M}(j_k^*)|}{\pi}$ , which is used for the kitchen datasets.

As shown in Fig. 7 and Table I, our proposed algorithms (skid-SLAM and skill-SLAM) are able to recover from several errors, such as missing landmarks and extra landmarks, made by kSLAM and reduce the trajectory error in both cases (i) and (ii). In addition, our proposed algorithms exhibit a clear speed advantage on these larger datasets due to the savings discussed in Section IV-C, continuing the trend observed in Section V-A. The more challenging layout in case (ii) does degrade performance, causing landmarks of the same class to be merged rather than correctly distinguished. However, the estimated landmark labels are always correct. The object-centric SLAM baseline [2] clearly performs worse. We find that it relies heavily on the manually set covariance matrices and thresholds.

### C. Pool Dataset with Real-Valued Feature Vectors

The kitchen dataset evaluates our algorithm with one-hot feature vectors from a closed-set object detector, where

TABLE I: Absolute trajectory error (m) and runtime (s) on the kitchen datasets.

| Dataset    | $N$  | $K_{gt}$ | $M$  | Odometry | Object-Centric [2] | kSLAM       | skid-SLAM   | skill-SLAM  |
|------------|------|----------|------|----------|--------------------|-------------|-------------|-------------|
| Diff Class | 3719 | 21       | 6552 | 0.264    | 0.361              | 0.034 / 429 | 0.033 / 325 | 0.033 / 364 |
| Same Class | 4009 | 21       | 6872 | 0.200    | 0.147              | 0.039 / 503 | 0.039 / 321 | 0.037 / 342 |

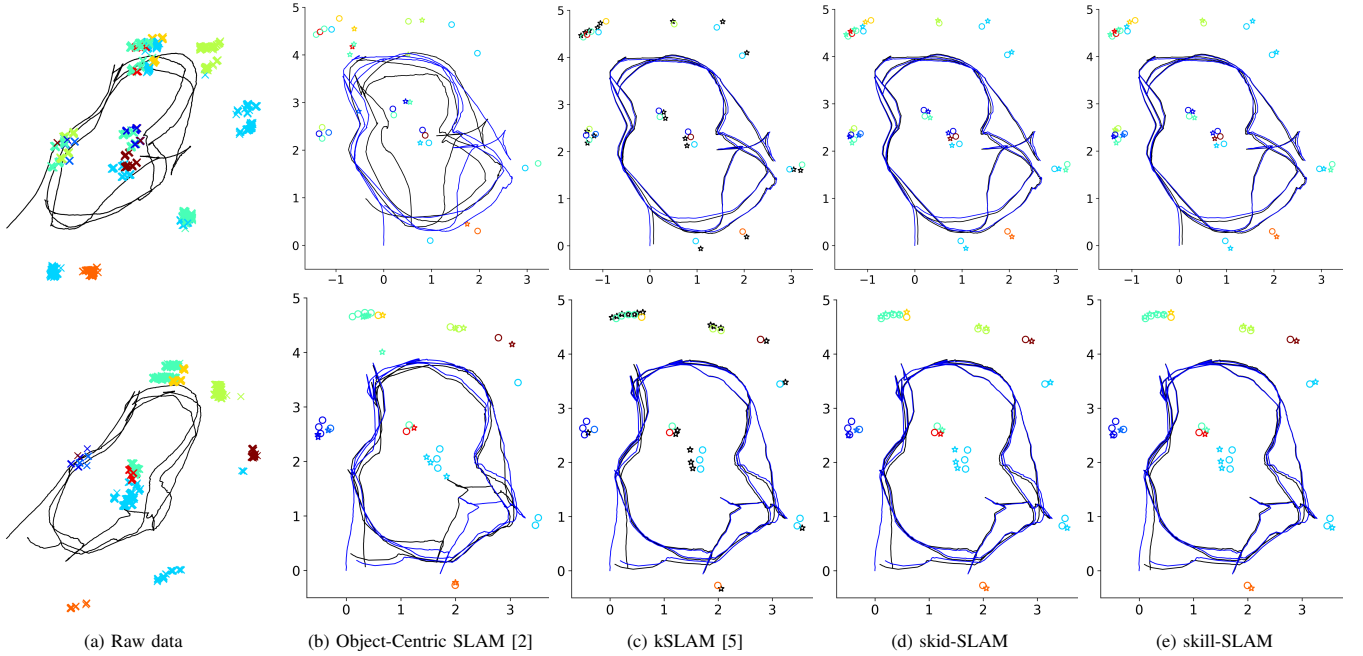


Fig. 7: Visualization of the kitchen results. Circles indicate pseudo-ground-truth landmark positions (which are slightly offset from the actual landmark positions). Crosses represent landmark measurements. Stars indicate estimated landmark positions. The colors of landmarks denote their class labels (if applicable). The blue and black trajectories are the pseudo-ground-truth and estimated trajectories, respectively. Top: Diff Class. Bottom: Same Class.

landmark classes are predefined and semantically distinct by construction. To evaluate our algorithms under the more challenging open-set setting, where object classes are unknown and real-valued feature vectors are required, we use the indoor pool dataset from [2]. This dataset was collected with a BlueROV2 equipped with a monocular camera and an Oculus m1200d multibeam sonar while navigating an indoor pool that contains diverse objects (e.g., lobster cages, seaweed, tires, etc.) (Fig. 8a). To represent the semantic identities of the objects, DINO features [21], extracted via unsupervised segmentation, are used instead of one-hot feature vectors, yielding real-valued descriptors for each detected object. Since ground-truth trajectory data is not available for this dataset, we evaluate the results through qualitative visualization of the estimated maps.

In Fig. 8, our proposed skill-SLAM is compared with the batch-processing kSLAM [5], but readers can also refer to [2] for additional comparison with their results. The heuristic in Section IV-A is used to set  $\beta$  since the landmarks in this dataset are well-separated.

In Fig. 8b, the raw data contains several errors, particularly for small and hard-to-see objects such as the hand net and the pipes. Both kSLAM and skill-SLAM merge these erroneous measurements into nearby landmarks. We hypothesize that the outer algorithm (3) decides not to introduce more landmarks for these sporadic outliers because the objective

value reduction in  $f_{slam}^*(K)$  for doing so does not outweigh the penalty  $\beta K$  in (3); thus, the inner algorithm (9) is forced to assign them incorrectly to nearby landmarks.

Owing to the noisy depth measurements from the sonar and accumulated odometry drift over time, when the robot loops back and re-observes a previous landmark (e.g., the seaweed and cage on the right side in Fig. 8b), its measurements projected from the raw odometry can be quite far away from its measurements in the first pass, making them difficult to associate correctly. In Fig. 8c, kSLAM erroneously merges the cage and seaweed together and makes similar mistakes elsewhere (e.g., between the cage and seaweed on the left side). However, skill-SLAM correctly associates the measurements of these objects by leveraging the semi-incremental scheme (Section IV-C) to reduce drift and semantic feature vectors to better distinguish between objects. We observe that the semantic information is critical here for data association, as the positional measurements are too noisy.

## VI. CONCLUSION

We extended the DAF-SLAM framework [5] and developed the skid-SLAM and skill-SLAM algorithms, with skill-SLAM performing slightly better than skid-SLAM. The extended framework and its algorithms additionally process and estimate semantic vectors of object landmarks, operate in

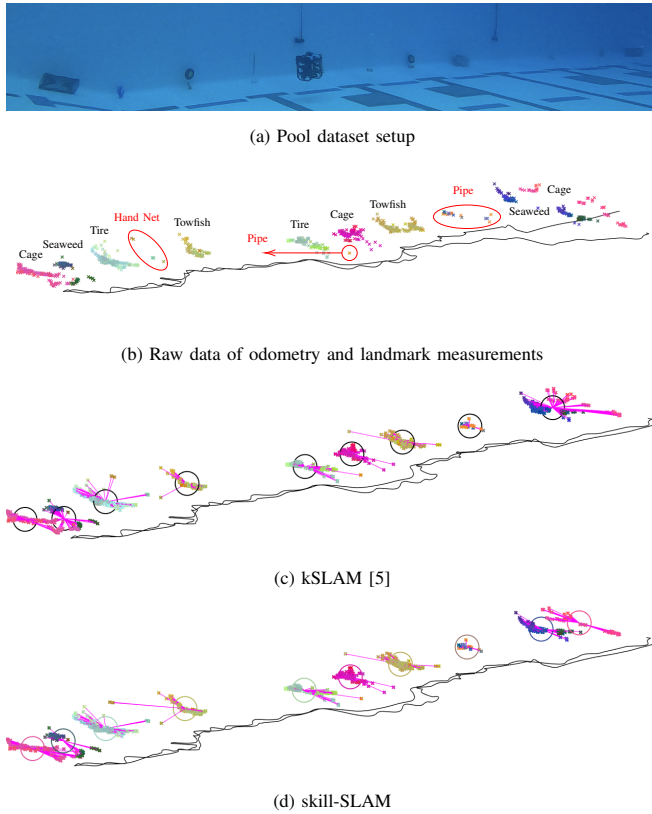


Fig. 8: Pool dataset results. The high-dimensional feature-vector measurements are reduced to three-dimensional vectors using PCA and mapped to RGB colors for visualization. The raw data has a few serious errors indicated by the red circles in Fig. 8b. In the kSLAM and skill-SLAM results, the magenta lines represent the estimated data associations, and the colored circles indicate the estimated landmark positions and semantics.

a semi-incremental manner to reduce drift, and provide principled guidelines and heuristics for setting a key parameter. They outperform the original kSLAM algorithm within the DAF-SLAM framework as well as several strong baselines on both synthetic and real-world datasets. An important future direction is handling both positional (e.g., spurious depths) and semantic (e.g., false positives) outlier measurements, potentially by modeling outliers in the generative models and using robust optimization objectives. Another future direction is incorporating shape estimation to enable a synergy between shape estimation and data association, and move beyond point-landmark representations.

## REFERENCES

- [1] K. J. Doherty, D. P. Baxter, E. Schneeweiss, and J. J. Leonard, "Probabilistic data association via mixture models for robust semantic SLAM," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1098–1104.
- [2] K. Singh, J. Hong, N. R. Rypkema, and J. J. Leonard, "Opti-acoustic semantic SLAM with unknown objects in underwater environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 1169–1176.
- [3] J. Hong, R. Choi, and J. J. Leonard, "Semantic enhancement for object slam with heterogeneous multimodal large language model agents," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 9036–9043.

- [4] E. Michael, T. Summers, T. A. Wood, C. Manzie, and I. Shames, "Probabilistic data association for semantic SLAM at scale," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4359–4364.
- [5] Y. Zhang, O. A. Severinsen, J. J. Leonard, L. Carlone, and K. Khosoussi, "Data-association-free landmark-based SLAM," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8349–8355.
- [6] J. Neira and J. D. Tardós, "Data association in stochastic mapping using the joint compatibility test," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 6, pp. 890–897, 2001.
- [7] M. Montemerlo, S. Thrun, D. Koller, B. Wegbreit et al., "FastSLAM: A factored solution to the simultaneous localization and mapping problem," *Aaai/iaai*, vol. 593598, no. 2, pp. 593–598, 2002.
- [8] K. J. Doherty, Z. Lu, K. Singh, and J. J. Leonard, "Discrete-continuous smoothing and mapping," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12395–12402, 2022.
- [9] V. Korotkine, M. Cohen, and J. R. Forbes, "Globally optimal data-association-free landmark-based localization using semidefinite relaxations," *IEEE Robotics and Automation Letters*, 2025.
- [10] B. Mu, S.-Y. Liu, L. Paull, J. Leonard, and J. P. How, "SLAM with objects using a nonparametric pose graph," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4602–4609.
- [11] J. Zhang, L. Yuan, T. Ran, Q. Tao, and L. He, "Bayesian nonparametric object association for semantic SLAM," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5493–5500, 2021.
- [12] T. Ran, L. Yuan, J. Zhang, L. He, R. Huang, and J. Mei, "Not only look but infer: Multiple hypothesis clustering of data association inference for semantic SLAM," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.
- [13] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic SLAM," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.
- [14] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "SLAM++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.
- [15] J. G. Rogers, A. J. Trevor, C. Nieto-Granda, and H. I. Christensen, "Simultaneous localization and mapping with learned object recognition and semantic data association," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1264–1270.
- [16] K. Li, D. DeTone, Y. F. S. Chen, M. Vo, I. Reid, H. Rezatofighi, C. Sweeney, J. Straub, and R. Newcombe, "ODAM: Object detection, association, and mapping using posed RGB video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5998–6008.
- [17] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 826–840, 2013.
- [18] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.
- [19] F. Dellaert and M. Kaess, "Georgia tech smoothing and mapping (gtsam)," 2019, 2019.
- [20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [21] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision Transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
- [22] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [23] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv preprint arXiv:1801.09847*, 2018.
- [24] F. Steinbrücker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense RGB-D images," in *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*. IEEE, 2011, pp. 719–722.
- [25] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 4193–4198.

## APPENDIX

### ASSUMPTIONS IN DERIVING THE UPPER BOUND FOR $\beta_{LB}$

Several assumptions used to reach (7) are briefly touched on in Section IV-A. We make them more explicit in this section.

First,  $K^*$  is defined to be the ground-truth number of landmarks. However, the optimal configuration  $f_{slam}^*(K^*)$  may not contain all ground-truth associations if our objective function (2) does not model the real world accurately. Assuming  $f_{slam}^*(K^*)$  contains all correct associations makes our reasoning much easier because we no longer need to consider the case where  $f_{slam}^*(K^*)$  already includes mis-associations. The validity of this assumption lies in the soundness of the modeling choices (i.e., Gaussian noise and the max-mixture model for data association) and is beyond the scope of this paper.

Second, the validity of (7) assumes that  $f_{slam}^*(K^* + 1)$  also does not contain incorrect associations. In other words, the extra landmark would only split the measurements associated with a landmark in  $f_{slam}^*(K^*)$ . This assumption is consistent with our working regime of large inter-landmark spacing relative to the landmark measurement noise, because any mis-association in this regime would introduce large residuals, which are avoided by the optimization. As illustrated in Fig. 9b, the added landmark splits the two measurements associated with the left landmark in Fig. 9a, which is possible under our assumption. However, the added landmark instead merges two measurements previously associated with two different landmarks in Fig. 9c, which is not possible under our assumption. When the inter-landmark spacing is small relative to the landmark measurement noise, this assumption starts to break down (Fig. 10), but this also falls outside of the working regime of our heuristic.

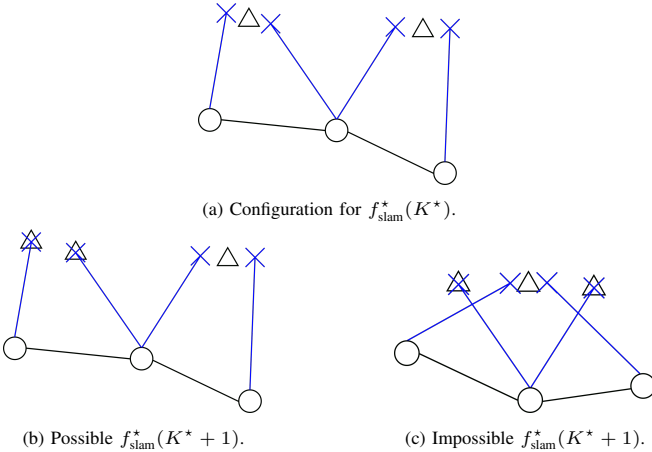


Fig. 9: Our assumption about the possible configuration for  $f_{slam}^*(K^* + 1)$ . The circles are robot poses. The triangles are landmarks. The crosses are measurements of landmarks.

Third, while not emphasized in the main paper, the odometry noise has to be moderate. In general, we find that increasing odometry noise has the opposite effect of increasing landmark measurement noise. In other words, it tends to increase  $\beta_{lb}$  and  $\beta_{ub}$  (Fig. 12) so  $\beta$  is more likely to fall

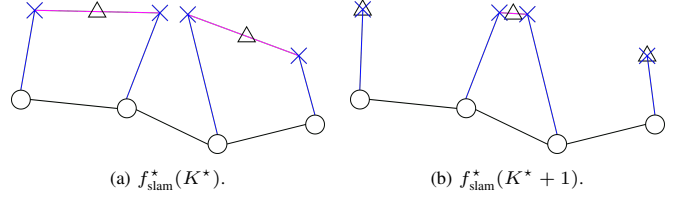


Fig. 10: An imaginary example where our assumption breaks. The magenta lines indicate data associations. The added landmark in 10b merges two measurements previously associated with two different landmarks.

below  $\beta_{lb}$ , causing overestimation of  $K$ . We hypothesize two possible mechanisms. (i) The SLAM optimization only cares about the relative weighting (or relative noise magnitude) between the odometry term and the landmark measurement term, because scaling the objective function (2) up or down does not change the result. Therefore, increasing odometry noise can be viewed as increasing the weighting of the landmark measurement term. The non-Gaussian residual increase and reduction for  $f_{slam}^*(K^* - 1)$  and  $f_{slam}^*(K^* + 1)$  relative to  $f_{slam}^*(K^*)$  are thus magnified, and  $\beta_{lb}$  and  $\beta_{ub}$  increase. (ii) Large odometry noise creates more trajectory drift, and landmark measurements projected from the trajectory (i.e., the robot poses) have greater dispersion than that implied by the nominal landmark measurement noise. Without known data associations, the larger spread shifts the residuals at each  $K$  in  $f_{slam}^*(K)$  upward.

Fourth, when adding a landmark, besides the direct residual reduction from the landmark measurement terms, there is a nonlinear relaxation effect on the residuals of the odometry measurements and other landmark measurements, as the trajectory becomes less constrained. We assume this effect is negligible compared to the drop in the landmark measurement residuals. In practice, we use a large probability value for our  $\chi^2$  heuristic so that  $\beta$  is considerably larger than the right-hand side of (7) to compensate for the unaccounted-for relaxation effect.

Under the assumptions made above, the best possible outcome of splitting the measurements associated with one landmark across two landmarks is to drive the residuals for all the measurements associated with that landmark to zero. The exception in Fig. 10 is a violation, since the residuals for all the measurements associated with the two landmarks are nearly driven to zero in Fig. 10b, which is even better than the outcome predicted by our upper bound (7).

### LOW INTER-LANDMARK SPACING SCENARIO

In Section IV-A, we mention that when the inter-landmark spacing is small relative to the landmark measurement noise,  $\beta$  is set empirically because several assumptions underlying the  $\beta$  heuristic no longer hold. However, useful guidelines and a reasonable starting point still exist. We discuss them in this section.

$\beta_{ub}$  is, in general, difficult to quantify since it depends on the environment-dependent mis-association errors. We thus focus on  $\beta_{lb}$ . It is easy to show, without additional

assumptions, that

$$\beta_{lb} \geq \max_k \|R_{i_k}^{*T} (y_{j_k}^* - t_{i_k}^*) - \bar{z}_k\|_{\Sigma}^2 \quad (11)$$

because the one extra landmark can always be associated with a single measurement to drive its residual to zero. The right-hand side of (11) follows a  $\chi^2$  distribution with degrees of freedom  $d$ . We can use the inverse  $\chi^2$  distribution and a large probability value to compute a conservative value for the right-hand side of (11), and  $\beta$  needs to be greater than this value according to (6).

As discussed in Section IV-A,  $\beta_{ub}$  is small in the regime of low inter-landmark spacing, and thus, we would like a value smaller than the right-hand side of (7) to set  $\beta$ . In this regime,  $f_{slam}^*(K^* + 1)$  may contain mis-associations such as the scenario in Fig. 10. However, it is still useful to consider the case where there are no mis-associations and the added landmark only splits the measurements associated with a landmark in  $f_{slam}^*(K^*)$ , which is essentially a restriction of the solution space of  $\min f_{slam}(K^* + 1)$ , and the resulting objective value will be larger than  $f_{slam}^*(K^* + 1)$ . While (7) gives an upper bound on the residual reduction for splitting the measurements, the goal here is to quantify the reduction more exactly to set a smaller  $\beta$ . We thus model splitting the measurements of a landmark as splitting Gaussian samples. When the number of samples is large (i.e.,  $|\mathcal{M}(j_k^*)| \rightarrow \infty$ ), the maximum reduction in the residuals (i.e., analogous to  $f_{slam}^*(K^*) - f_{slam}^*(K^* + 1)$ ) is approximately  $\frac{2|\mathcal{M}(j_k^*)|}{\pi}$ , which can be a starting point for tuning  $\beta$ . We leave the derivations in the final section.

We summarize the findings here: (i)  $\beta$  should be chosen to be greater than the right-hand side of (11). (ii) In the modeling above, we have restricted the solution space, and a lower  $f_{slam}^*(K^* + 1)$  may actually be achievable. In other words,  $\beta_{lb} = f_{slam}^*(K^*) - f_{slam}^*(K^* + 1)$  may actually be greater than  $\frac{2|\mathcal{M}(j_k^*)|}{\pi}$ , and setting  $\beta$  to this value may cause  $\beta$  to fall below  $\beta_{lb}$ . (iii) The relaxation effect of adding a landmark on the odometry terms and other measurement terms is not considered. In principle, the relaxation effect would make  $f_{slam}^*(K^* + 1)$  even smaller, which may cause  $\beta = \frac{2|\mathcal{M}(j_k^*)|}{\pi}$  to fall below  $\beta_{lb}$ . (iv) The result assumes a large number of samples (measurements of a landmark). This is typically true with a high-frame-rate sensor. When this assumption is violated, a lower  $f_{slam}^*(K^* + 1)$  may be achievable, and  $\beta = \frac{2|\mathcal{M}(j_k^*)|}{\pi}$  may fall below  $\beta_{lb}$ .

#### APPLICABLE RANGES OF THE $\beta$ HEURISTICS

It is interesting to examine what the inverse  $\chi^2$  heuristic in Section IV-A and setting  $\beta = \frac{2|\mathcal{M}(j_k^*)|}{\pi}$  lead to under different inter-landmark spacing (relative to the landmark measurement noise). To this end, we first summarize the properties of our datasets (Section V) in Tables II and III. The kitchen datasets (i.e., Diff Class and Same Class) clearly have their measurement standard deviation  $\sigma$  on the same order of magnitude as the mean and minimum inter-landmark spacing, whereas the other datasets have  $\sigma$  orders of magnitude smaller than the spacing. Accordingly, we

use  $\beta = \frac{2|\mathcal{M}(j_k^*)|}{\pi}$  for the kitchen datasets (see the actual parameters used in Table V).

We also perform a set of small-scale experiments for simple validation in Table IV. In Experiments 1 – 3, we see that  $\beta$  has to scale with the number of measurements per landmark ( $|\mathcal{M}(j)|$ ) to estimate  $K$  correctly. The  $\frac{2|\mathcal{M}(j)|}{\pi}$  heuristic consistently sets  $\beta$  too small, leading to overestimation of  $K$  in Experiments 4 – 6, which is expected because our assumptions used to derive the heuristic tend to underestimate what  $\beta$  should ideally be (as explained in the section “Low Inter-Landmark Spacing Scenario”).

In Experiments 7 – 14, as  $\sigma$  approaches the same order of magnitude as the mean inter-landmark spacing, the inverse  $\chi^2$  heuristic starts to underestimate  $K$  because of the drop in  $\beta_{ub}$  (Section IV-A), while the  $\frac{2|\mathcal{M}(j)|}{\pi}$  heuristic begins to estimate  $K$  correctly. In Experiments 10 and 14,  $\sigma = 1.5$  is considered extremely large, but the  $\frac{2|\mathcal{M}(j)|}{\pi}$  heuristic manages to correctly estimate  $K$ .

#### ESTIMATING THE NUMBER OF MEASUREMENTS PER LANDMARK

In Section IV-A, to compute a heuristic for setting  $\beta$ , we need to estimate the number of measurements per landmark. We consider a forward-moving model (Fig. 11a) for the kitchen dataset (Section V-B) and a sideward-moving model (Fig. 11b) for the pool dataset (Section V-C). These are simple 2D models used to simplify the calculations.

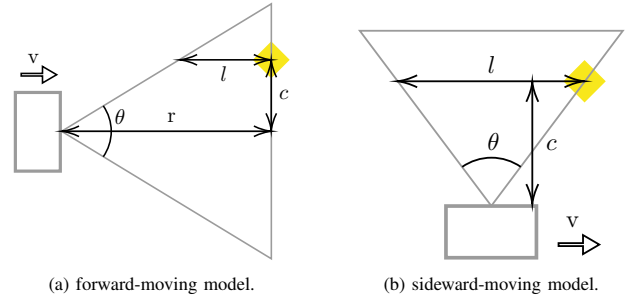


Fig. 11: Simple 2D models to estimate the number of measurements per landmark. The sensor is represented as a gray rectangle. The yellow diamond represents a landmark.

In Fig. 11a, the distance ( $l$ ) while the landmark is in view, is

$$l = r - c \cot\left(\frac{\theta}{2}\right) \quad (12)$$

where  $r$  is the cutoff depth,  $\theta$  is the field of view, and  $c$  may be computed as an average over all landmark measurements. In Fig. 11b, the in-view distance ( $l$ ) is

$$l = 2c \tan\left(\frac{\theta}{2}\right) \quad (13)$$

where  $c$  is also computed as an average over all landmark measurements, with  $c$  corresponding to the depth component here. The estimated number of observations is then computed as  $\text{fps} \times l/v$  where  $\text{fps}$  is the sensor fps and  $v$  is the robot's average speed computed from odometry measurements. These models assume that the robot passes by each landmark only once, which is generally true for the segment processing

TABLE II: Dataset properties to the best of our estimates.

| Dataset    | $N$  | $K_{\text{gt}}$      | $M$  | $ \mathcal{M}(j) _{\text{avg}}$ | $\sigma$                   | Spacing <sup>1</sup> (mean, min, max) |
|------------|------|----------------------|------|---------------------------------|----------------------------|---------------------------------------|
| Diff Class | 3719 | 21                   | 6552 | 312                             | 0.1                        | 0.38, 0.10, 1.44                      |
| Same Class | 4009 | 21                   | 6872 | 327                             | 0.1                        | 0.31, 0.11, 1.88                      |
| Pool       | 2335 | 12 (10) <sup>2</sup> | 1104 | 92 (110) <sup>2</sup>           | 0.3, 0.3, 0.6 <sup>3</sup> | 1.74, 1.17, 2.94                      |
| Grid 3D    | 216  | 43                   | 430  | 10                              | 0.05                       | 1.06, 0.49, 2.00 <sup>4</sup>         |
| Grid 2D    | 500  | 100                  | 1000 | 10                              | 0.05                       | 1.33, 0.18, 3.51 <sup>4</sup>         |

<sup>1</sup> Defined as the distance to the closest landmark for every landmark.

<sup>2</sup> Two landmarks are almost entirely missing in the measurements (Fig. 8b).

<sup>3</sup> In the order of x, y and z axes in the robot camera frame.

<sup>4</sup> Example numbers from one simulation instance.

TABLE III: Additional dataset properties to the best of our estimates.

| Dataset    | Avg. Odom. Std. x y z r p y ( $\times 10^{-3}$ ) | Trajectory Size <sup>1</sup> | Mislabeled Rate <sup>2</sup> | Semantic Std. |
|------------|--------------------------------------------------|------------------------------|------------------------------|---------------|
| Diff Class | 2.5, 1.6, 1.8, 3.6, 4.7, 2.3                     | $3.6 \times 4.6$             | -                            | -             |
| Same Class | 2.6, 1.7, 1.8, 3.6, 4.7, 2.2                     | $3.4 \times 4.6$             | -                            | -             |
| Pool       | 1.0, 1.0, 1.0, 1.0, 1.0, 1.0                     | $20.0 \times 1.9$            | -                            | -             |
| Grid 3D    | 50, 50, 50, 5.0, 5.0, 5.0                        | $6 \times 6 \times 6$        | 0.1                          | 0.1           |
| Grid 2D    | 50, 50, 5.0                                      | $20 \times 25$               | 0.1                          | 0.1           |

<sup>1</sup> If the trajectory is mostly on a principal plane, this is the span of the trajectory on the plane.

<sup>2</sup> Nominal mislabeling rate to simulate an object detector.

TABLE IV: Results using the two  $\beta$  heuristics on a small-scale  $3 \times 2 \times 3$  simulated 3D Grid dataset (18 poses).

| Exp. | $K_{\text{gt}}$ | $ \mathcal{M}(j) $ | $\sigma$ | Spacing <sup>1</sup> (mean, min, max) | Heuristic                    | $\beta$ | $K_{\text{est}}$ |
|------|-----------------|--------------------|----------|---------------------------------------|------------------------------|---------|------------------|
| 1    | 3               | 100                | 0.05     | 1.26, 1.16, 1.47                      | $(\chi^2)^{-1}(0.999, 300)$  | 381.4   | 3 (0)            |
| 2    | 3               | 500                | 0.05     | 0.94, 0.43, 1.96                      | $(\chi^2)^{-1}(0.999, 1500)$ | 1675    | 3 (0)            |
| 3    | 3               | 800                | 0.05     | 1.03, 0.91, 1.26                      | $(\chi^2)^{-1}(0.999, 2400)$ | 2620    | 3 (0)            |
| 4    | 3               | 100                | 0.05     | 1.26, 1.16, 1.47                      | $\frac{2 \times 100}{\pi}$   | 63.7    | 6 (+3)           |
| 5    | 3               | 500                | 0.05     | 0.94, 0.43, 1.96                      | $\frac{2 \times 500}{\pi}$   | 318.3   | 4 (+1)           |
| 6    | 3               | 800                | 0.05     | 1.03, 0.91, 1.26                      | $\frac{2 \times 800}{\pi}$   | 509.3   | 4 (+1)           |
| 7    | 10              | 500                | 0.1      | 0.57, 0.22, 1.18                      | $(\chi^2)^{-1}(0.999, 1500)$ | 1675    | 9 (-1)           |
| 8    | 10              | 500                | 0.5      | 0.57, 0.24, 1.22                      | $(\chi^2)^{-1}(0.999, 1500)$ | 1675    | 4 (-6)           |
| 9    | 10              | 500                | 0.8      | 0.57, 0.25, 1.24                      | $(\chi^2)^{-1}(0.999, 1500)$ | 1675    | 4 (-6)           |
| 10   | 10              | 500                | 1.5      | 0.48, 0.26, 1.11                      | $(\chi^2)^{-1}(0.999, 1500)$ | 1675    | 3 (-7)           |
| 11   | 10              | 500                | 0.1      | 0.57, 0.22, 1.18                      | $\frac{2 \times 500}{\pi}$   | 318.3   | 18 (+8)          |
| 12   | 10              | 500                | 0.5      | 0.57, 0.24, 1.22                      | $\frac{2 \times 500}{\pi}$   | 318.3   | 14 (+4)          |
| 13   | 10              | 500                | 0.8      | 0.57, 0.25, 1.24                      | $\frac{2 \times 500}{\pi}$   | 318.3   | 11 (+1)          |
| 14   | 10              | 500                | 1.5      | 0.48, 0.26, 1.11                      | $\frac{2 \times 500}{\pi}$   | 318.3   | 10 (0)           |

<sup>1</sup> Distance to the closest landmark for every landmark, calculated from the estimated landmark positions given the ground truth data associations, to reflect the configuration of  $f_{\text{slam}}^*(K^*)$ .

stage (Section IV-C). If there is a strong belief that the robot passes by each landmark multiple times, we can set a different  $\beta$  for the whole-trajectory global optimization. Otherwise, we rely on the robustness of formulation (3) to correctly estimate  $K$ . For the segment processing, the number of observations should also not exceed the number of poses in the segment, assuming each pose can produce at most one observation of a landmark.

#### ALGORITHM PARAMETERS

We provide our parameter settings in Table V.

#### LANDMARK EVALUATION

We provide additional details about the evaluation of landmark estimation on the synthetic datasets. The data association is unknown to our algorithms, and therefore, we need to match the estimated landmarks to the reference landmarks. The reference landmarks are obtained along with

the reference trajectory, by optimizing the measurements given the ground-truth data associations. We first compute the SIM(3) alignment between the estimated trajectory and the reference trajectory, and apply it to the estimated landmark positions. A linear assignment problem is then solved to match the estimated landmark positions to reference landmark positions. Only  $\min(K_{\text{ref}}, K_{\text{est}})$  pairs are matched, and any extra landmarks are not included in the evaluation. The estimation error is computed over all matched pairs.

#### VISUALIZATION OF $f_{\text{SLAM}}^*(K)$

It is interesting to plot  $f_{\text{slam}}^*(K)$  as a function of  $K$ . Computing the true  $f_{\text{slam}}^*(K)$  is combinatorial, therefore we use our inner algorithm (2) to approximate its value. We show three cases under different noise levels in Fig. 12. In the case of the normal noise level, we can clearly observe the elbow point at the true  $K$ , which is consistent with our discussions in Section IV-A. The case of high

TABLE V: Motion model parameters (calculated based on prior information or from measurements) and algorithm parameters.

| Dataset  | $v$   | $\theta$ | $r$ | $c$   | fps  | $ \mathcal{M}(j) _{\text{est}}$ | Heuristic                   | $\beta_{\text{glb}}$ | $\frac{\sigma}{\sigma_s} w_s^1$ | $N_n$ | $\beta_{\text{seg}}^2$ | $N_{\text{inner}}^3$ | $N_{\text{LM}}^4$ | k-means <sup>5</sup> |
|----------|-------|----------|-----|-------|------|---------------------------------|-----------------------------|----------------------|---------------------------------|-------|------------------------|----------------------|-------------------|----------------------|
| Grid 3D  | -     | -        | -   | -     | -    | 10                              | $(\chi^2)^{-1}(0.999, 30)$  | 60                   | 0.1                             | 100   | 60                     | 30                   | 100               | 3, 15, 0.1           |
| Grid 2D  | -     | -        | -   | -     | -    | 10                              | $(\chi^2)^{-1}(0.999, 20)$  | 45                   | 0.1                             | 100   | 45                     | 30                   | 100               | 3, 15, 0.1           |
| Diff Cls | 0.247 | 120      | 3   | 0.642 | 20   | 213                             | $\frac{2 \times 213}{\pi}$  | 136                  | 0.5                             | 100   | 64                     | 15                   | 100               | 3, 15, 0.1           |
| Same Cls | 0.246 | 120      | 3   | 0.698 | 20   | 211                             | $\frac{2 \times 211}{\pi}$  | 134                  | 0.5                             | 100   | 64                     | 15                   | 100               | 3, 15, 0.1           |
| Pool     | 0.186 | 80       | -   | 2.62  | 3.81 | 90                              | $(\chi^2)^{-1}(0.999, 270)$ | 347                  | 100                             | 100   | 347                    | 15                   | 100               | 3, 15, 0.1           |

<sup>1</sup> Real-valued feature vectors are additionally normalized by length.

<sup>2</sup> Calculated using  $N_n$  instead of  $|\mathcal{M}(j)|_{\text{est}}$  if  $N_n < |\mathcal{M}(j)|_{\text{est}}$ .

<sup>3</sup> Number of inner solver iterations. The argmin solution across iterations (not necessarily the final iteration due to k-means++ randomness) is selected.

<sup>4</sup> Number of Levenberg-Marquardt optimizer iterations for SLAM.

<sup>5</sup> Number of k-means attempts (the argmin result is kept), termination iteration and termination tolerance.

odometry noise exhibits the opposite trend compared to the case of high landmark measurement noise. We provide two possible explanations in the previous section “Assumptions in Deriving the Upper Bound for  $\beta_{\text{lb}}$ ”. Additionally, the case of high odometry noise appears to be “noisier”, which reveals that the algorithm more easily returns a suboptimal solution, potentially due to the drift accumulation effect of odometry noise that does not occur with landmark measurement noise.

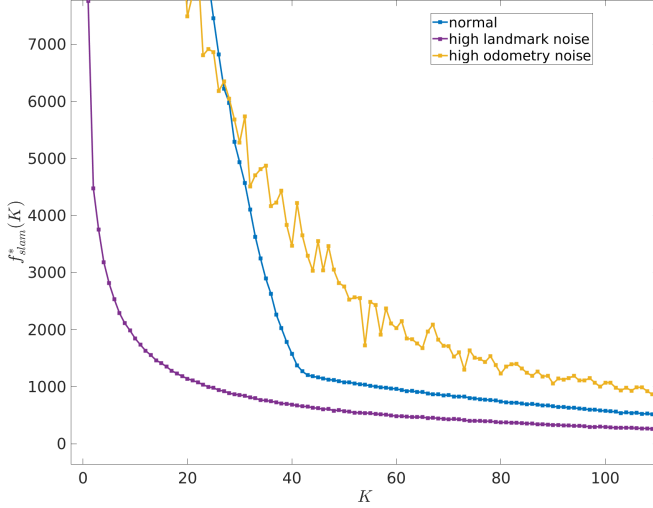


Fig. 12: Evaluating  $f_{\text{slam}}^*(K)$  using the inner algorithm (2) on the synthetic 3D grid dataset of size  $6 \times 6 \times 6$ . The normal case uses a landmark measurement standard deviation 0.1 and odometry translation and rotation standard deviations of 0.05 and 0.005, respectively. The high landmark noise case uses a landmark measurement standard deviation 0.5, which is fairly large with respect to the map scale. The high odometry noise case uses translation and rotation standard deviations of 0.15 and 0.015, respectively. The true  $K$  is 43.

There are two effects that lead to the differences between the normal case and the case of high landmark measurement noise. First, the nonlinear optimization will result in a lower SLAM objective value as the landmark measurement noise increases. We show a simple experiment in Table VI for a standard landmark-based SLAM problem (given data associations). The data are generated synthetically from the Gaussian noise models that exactly match the probabilistic models assumed by the SLAM objective function. Since the covariance matrices and the probabilistic models are exact, the covariance matrices perfectly normalize the measurement residuals (1), and thus, given that the ground truth is used as

the initial values for the optimization variables, the initial errors (i.e., objective function values) are all the same. However, after optimization, due to the nonlinear interplay between the odometry term and the landmark measurement term, the optimal objective values are smaller at higher landmark measurement noise. This explains why the case of high landmark measurement noise has lower  $f_{\text{slam}}^*(K)$  than the normal case at the true  $K$  in Fig. 12.

TABLE VI: Factor graph errors (i.e. objective function values) of a standard landmark-based SLAM problem (given data associations) at different landmark measurement noise levels. The synthetic 3D grid dataset is used. Landmark measurements are generated from the Gaussian generative model (1), so the covariance matrices are exact. Ground truth is used as the initial guess.

| LM Std        | 0.001 | 0.01 | 0.05 | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  |
|---------------|-------|------|------|------|------|------|------|------|
| Initial Error | 2607  |      |      |      |      |      |      |      |
| Final Error   | 1230  | 1222 | 1219 | 1203 | 1183 | 1174 | 1169 | 1167 |

Second, when the landmark measurement noise is high (relative to the inter-landmark spacing), the unnormalized (i.e., not normalized by the covariance matrix) residual  $\|R_{i_k}^T(y_{j_k} - t_{i_k}) - \tilde{z}_k\|^2$  of an association error may have a similar magnitude as the residual due to pure sensor noise. Therefore, even when  $K$  is below the true  $K$  and more association errors are introduced in computing  $f_{\text{slam}}^*(K)$ , we do not observe as sharp a transition as in the normal case in Fig. 12. This is where the gap between  $\beta_{\text{lb}}$  and  $\beta_{\text{ub}}$  becomes narrow, and our algorithm (3), using  $\beta$  set by the inverse  $\chi^2$  heuristic in (7), begins to underestimate  $K$ .

#### KITCHEN DATASET PANORAMAS

For direct visualization of the kitchen datasets, we provide panoramas of the setups in Fig. 13.

#### SPLITTING GAUSSIAN SAMPLES

Following up on the previous section “Low Inter-Landmark Spacing Scenario”, we model the addition of a landmark as splitting Gaussian samples to approximate the objective value change from  $f_{\text{slam}}^*(K^*)$  to the pre-optimization (non-optimal) value of  $f_{\text{slam}}^*(K^* + 1)$ . Fig. 14 shows a schematic for illustration. For clarity, we hereafter use notation that is independent of the rest of the paper.



(a) Kitchen - Diff Class.



(b) Kitchen - Same Class.

Fig. 13: Panoramas of the kitchen environment and the experiment setups. In the datasets, we have removed the laptop detections, because the laptop screens caused erroneous depth measurements.



Fig. 14: Splitting Gaussian samples to model  $f_{\text{slam}}(K^* + 1)$ . We show two different ways of splitting.

#### A. Problem Setup

Consider a landmark at position  $\mu \in \mathbb{R}^d$  observed  $n$  times. Let  $u_1, \dots, u_n \in \mathbb{R}^d$  denote the measurements for this landmark. We assume an isotropic noise covariance. The anisotropic case can be derived from the isotropic case with additional tools such as PCA. Further, without loss of generality, we assume that the noise standard deviation is unity; otherwise, the residuals in our derivations below would be scaled by the inverse variance throughout (since SLAM optimization assumes a fixed covariance), but the conclusion in Subsection C remains unchanged due to cancellation. In short,

$$u_1, \dots, u_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, I_d).$$

In this model, we use the average as our estimate,

$$\bar{u} = \frac{1}{n} \sum_{i=1}^n u_i.$$

The total residual (after normalizing by the variance) for all the measurements is

$$r_0 = \sum_{i=1}^n \|u_i - \bar{u}\|^2, \quad (14)$$

which is analogous to the total residual for all measurements of a landmark in SLAM.

#### B. Reduction Formula for Any Split

Suppose the measurements are split into two nonempty groups  $G_1$  and  $G_2$  with sizes  $n_1 = |G_1|$  and  $n_2 = |G_2|$  ( $n_1 + n_2 = n$ ). Each has its own average estimate,

$$\bar{u}_1 = \frac{1}{n_1} \sum_{i \in G_1} u_i, \quad \bar{u}_2 = \frac{1}{n_2} \sum_{i \in G_2} u_i.$$

The total residual after splitting is now

$$r_{\text{split}} = \sum_{i \in G_1} \|u_i - \bar{u}_1\|^2 + \sum_{i \in G_2} \|u_i - \bar{u}_2\|^2. \quad (15)$$

*Lemma 1 (Residual Reduction):*

For any split  $G_1$  and  $G_2$ ,

$$r_0 - r_{\text{split}} = n_1 \|\bar{u}_1 - \bar{u}\|^2 + n_2 \|\bar{u}_2 - \bar{u}\|^2.$$

Equivalently,

$$r_0 - r_{\text{split}} = \frac{n_1 n_2}{n} \|\bar{u}_1 - \bar{u}_2\|^2. \quad (16)$$

*Proof:*

Expanding  $\sum_{i \in G_1} \|u_i - \bar{u}_1\|^2$ , we obtain

$$\begin{aligned} & \sum_{i \in G_1} \|u_i\|^2 + \sum_{i \in G_1} \|\bar{u}_1\|^2 - 2 \sum_{i \in G_1} u_i^\top \bar{u}_1 \\ &= \sum_{i \in G_1} \|u_i\|^2 + n_1 \|\bar{u}_1\|^2 - 2n_1 \|\bar{u}_1\|^2 \\ &= \sum_{i \in G_1} \|u_i\|^2 - n_1 \|\bar{u}_1\|^2, \end{aligned}$$

and similarly for  $\sum_{i \in G_2} \|u_i - \bar{u}_2\|^2$ . After merging terms,

$$\begin{aligned} r_{\text{split}} &= \sum_{i=1}^n \|u_i\|^2 - n_1 \|\bar{u}_1\|^2 - n_2 \|\bar{u}_2\|^2 \\ &= \sum_{i=1}^n \|u_i - \bar{u}\|^2 - n_1 \|\bar{u}_1\|^2 - n_2 \|\bar{u}_2\|^2 \\ &\quad - n \|\bar{u}\|^2 + 2n \|\bar{u}\|^2 \\ &= r_0 + n_1 (\|\bar{u}_1\|^2 - \|\bar{u}\|^2) + n_2 (\|\bar{u}_2\|^2 - \|\bar{u}\|^2). \end{aligned}$$

Rearranging terms,

$$r_0 - r_{\text{split}} = n_1 (\|\bar{u}_1\|^2 - \|\bar{u}\|^2) + n_2 (\|\bar{u}_2\|^2 - \|\bar{u}\|^2). \quad (17)$$

For the two terms in Lemma 1,

$$\begin{aligned} n_1 \|\bar{u}_1 - \bar{u}\|^2 &= n_1 \|\bar{u}_1\|^2 + n_1 \|\bar{u}\|^2 - 2n_1 \bar{u}_1^\top \bar{u}, \\ n_2 \|\bar{u}_2 - \bar{u}\|^2 &= n_2 \|\bar{u}_2\|^2 + n_2 \|\bar{u}\|^2 - 2n_2 \bar{u}_2^\top \bar{u}. \end{aligned}$$

Note the following equivalence,

$$2n_1 \bar{u}_1^\top \bar{u} + 2n_2 \bar{u}_2^\top \bar{u} = 2 \sum_{i=1}^n u_i^\top \bar{u} = 2(n_1 + n_2) \|\bar{u}\|^2,$$

which makes  $r_0 - r_{\text{split}}$  in Lemma 1 the same as (17),

$$n_1 (\|\bar{u}_1\|^2 - \|\bar{u}\|^2) + n_2 (\|\bar{u}_2\|^2 - \|\bar{u}\|^2)$$

This completes half of the proof.

For the equivalence part in Lemma 1, note the following,

$$\bar{u} = \frac{n_1 \bar{u}_1 + n_2 \bar{u}_2}{n},$$

plugging which into (17) leads to

$$\begin{aligned} r_0 - r_{\text{split}} &= n_1 \|\bar{u}_1\|^2 + n_2 \|\bar{u}_2\|^2 - \frac{\|n_1 \bar{u}_1 + n_2 \bar{u}_2\|^2}{n} \\ &= \frac{(nn_1 - n_1^2)}{n} \|\bar{u}_1\|^2 + \frac{(nn_2 - n_2^2)}{n} \|\bar{u}_2\|^2 - 2 \frac{n_1 n_2}{n} \bar{u}_1^\top \bar{u}_2 \end{aligned}$$

Noting that  $nn_1 - n_1^2 = n_1(n - n_1) = n_1 n_2$  and similarly for  $n_2$ , we can easily show that

$$\begin{aligned} r_0 - r_{\text{split}} &= \frac{n_1 n_2}{n} (\|\bar{u}_1\|^2 + \|\bar{u}_2\|^2 - 2\bar{u}_1^\top \bar{u}_2) \\ &= \frac{n_1 n_2}{n} \|\bar{u}_1 - \bar{u}_2\|^2, \end{aligned}$$

which completes the proof.  $\blacksquare$

We remark that  $r_0 - r_{\text{split}}$  is analogous to  $f_{\text{slam}}^*(K^*) - f_{\text{slam}}(K^* + 1)$ , where the second term is not necessarily optimal since SLAM optimization would distribute some residuals into the odometry term and other measurement terms so that the projected landmark positions could shift. Our splitting model is analogous to the pre-optimized value  $f_{\text{slam}}(K^* + 1)$ .

### C. Geometric Half-Space Split

We now consider a geometric split that divides the measurements into two halves, which is similar to the right schematic in Fig. 14. We start by choosing a unit vector

$$v \in \mathbb{R}^d, \quad \|v\| = 1.$$

Define the scalar projection

$$s_i = v^\top u_i.$$

Without loss of generality, we center our measurements

$$u_i \sim \mathcal{N}(0, I_d),$$

which leads to

$$s_i \sim \mathcal{N}(0, 1).$$

Split the samples according to the sign of this projection:

$$G_+ = \{i : s_i \geq 0\}, \quad G_- = \{i : s_i < 0\}.$$

An important assumption is that the number of measurements is large, and therefore,  $\bar{u} \approx \mu = 0$ . Because of this assumption, additionally we also have  $(I - vv^\top)\bar{u}_+ \approx (I - vv^\top)\bar{u}_- \approx (I - vv^\top)\bar{u} \approx 0$ . This implies that the residual reduction  $r_0 - r_{\text{split}}$  is only contributed by the component along the  $v$  direction, since the sum of the residuals in all orthogonal directions is unchanged from  $r_0$  (14) to  $r_{\text{split}}$  (15). Furthermore, due to this assumption, we have  $\bar{s}_+ = \frac{1}{n_+} \sum_{i \in G_+} s_i \approx \mathbb{E}[s \mid s \geq 0]$  and  $\bar{s}_- = \frac{1}{n_-} \sum_{i \in G_-} s_i \approx \mathbb{E}[s \mid s < 0]$ , where  $s$  follows  $\mathcal{N}(0, 1)$  and  $s = v^\top u$  for  $u$  following  $\mathcal{N}(0, I_d)$ .

More formally, consider  $u$ , which follows  $\mathcal{N}(0, I_d)$ , decomposed into the component along  $v$  and the component orthogonal to  $v$ :

$$u = sv + z, \quad z \perp v.$$

Note that  $z = (I - vv^\top)u$  and  $s = v^\top u$  are jointly Gaussian. For this zero-mean isotropic Gaussian case, it can be shown that the covariance between  $z$  and  $s$  is zero,

$$\begin{aligned} \text{COV}(s, z) &= \mathbb{E}[sz] - \mathbb{E}[s]\mathbb{E}[z] = \mathbb{E}[v^\top u(I - vv^\top)u] \\ &= \mathbb{E}[(I - vv^\top)uu^\top v] = (I - vv^\top)\mathbb{E}[uu^\top]v \\ &= (I - vv^\top)(\text{COV}(u) + \mathbb{E}[u]\mathbb{E}[u]^\top)v \\ &= v - vv^\top v = v - v\|v\|^2 = 0. \end{aligned}$$

For jointly Gaussian variables,  $z$  is then independent of  $s$ , which makes

$$\mathbb{E}[z \mid s \geq 0] = \mathbb{E}[z \mid s < 0] = \mathbb{E}[z] = 0.$$

Therefore,

$$\mathbb{E}[u \mid s \geq 0] = \mathbb{E}[sv + z \mid s \geq 0] = \mathbb{E}[s \mid s \geq 0]v,$$

and similarly,

$$\mathbb{E}[u \mid s < 0] = \mathbb{E}[s \mid s < 0]v.$$

By definition,

$$\mathbb{E}[s \mid s \geq 0] = \frac{\int_0^\infty s \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds}{\mathbb{P}(s \geq 0)}.$$

Since

$$\mathbb{P}(s \geq 0) = \frac{1}{2},$$

we get

$$\mathbb{E}[s \mid s \geq 0] = 2 \int_0^\infty s \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds.$$

Let

$$q = \frac{s^2}{2}, \quad dq = s ds.$$

Then

$$\mathbb{E}[s \mid s \geq 0] = \frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-q} dq.$$

Since

$$\int_0^\infty e^{-q} dq = 1,$$

we obtain

$$\mathbb{E}[s \mid s \geq 0] = \sqrt{\frac{2}{\pi}}.$$

By symmetry,

$$\mathbb{E}[s \mid s < 0] = -\sqrt{\frac{2}{\pi}}.$$

Therefore,

$$\mathbb{E}[u \mid s \geq 0] = \sqrt{\frac{2}{\pi}} v,$$

and

$$\mathbb{E}[u \mid s < 0] = -\sqrt{\frac{2}{\pi}} v.$$

Due to the assumption of a large number of measurements,

$$\bar{u}_+ \approx \mathbb{E}[u \mid s \geq 0] = \sqrt{\frac{2}{\pi}} v,$$

and

$$\bar{u}_- \approx \mathbb{E}[u \mid s < 0] = -\sqrt{\frac{2}{\pi}} v,$$

Also due to the large-sample assumption and the symmetry of an isotropic Gaussian, approximately half of the samples fall into each group:

$$n_+ \approx \frac{n}{2}, \quad n_- \approx \frac{n}{2}.$$

We substitute these results into Lemma 1 and obtain

$$\begin{aligned} r_0 - r_{\text{split}} &= \frac{n_+ n_-}{n} \|\bar{u}_+ - \bar{u}_-\|^2 \\ &= \frac{n}{4} \left\| \sqrt{\frac{2}{\pi}} v + \sqrt{\frac{2}{\pi}} v \right\|^2 \\ &= \frac{2n}{\pi} \|v\|^2 \\ &= \frac{2n}{\pi}. \end{aligned}$$

#### D. Optimality of the Half-Space Split

When the number of measurements  $n$  is large ( $n \rightarrow \infty$ ), the half-space split maximally reduces the residual (i.e., maximizes  $r_0 - r_{\text{split}}$ ). Intuitively, in (16),  $n_1 n_2 = n_1(n - n_1)$  is maximized (simply by setting the derivative to zero) when  $n_1 = n_2 = \frac{n}{2}$ , which is achieved by the half-space split when  $n$  is large. When moving away from the center split, the decrease in  $n_1 n_2$  dominates the growth in  $\|\bar{u}_1 - \bar{u}_2\|^2$  (one can reason about  $n_1$  and  $n_2$  through Gaussian CDF and about  $\bar{u}_1$  and  $\bar{u}_2$  using the previous derivations but with a non-zero threshold).

Under non-ideal conditions such as  $n \not\rightarrow \infty$  and an anisotropic data covariance (Fig. 15 for illustration),  $v$  may not be an arbitrary direction, but the principal direction of maximum data variance. Splitting at  $s = 0$  may not be optimal since the best split may shift toward outliers to maximize  $\|\bar{u}_1 - \bar{u}_2\|^2$ . The directions orthogonal to  $v$  may also contribute to the residual reduction under non-ideal conditions.



Fig. 15: Splitting Gaussian samples under non-ideal conditions.

We now provide a more thorough derivation of the optimality of the half-space split under ideal conditions. Our objective is to maximize (16), which is valid for all splits. For convenience, we repeat it below:

$$r_0 - r_{\text{split}} = \frac{n_1 n_2}{n} \|\bar{u}_1 - \bar{u}_2\|^2.$$

There are two parts: (i) showing that splitting at  $s = 0$  is optimal among all the half-space splits, and (ii) showing that the half-space splits are optimal among all splits. For the first part, we consider a general half-space split at  $s = t$  instead of at  $s = 0$ :

$$G_1 = \{i : s_i < t\}, \quad G_2 = \{i : s_i \geq t\}.$$

The objective is to show that  $t = 0$  maximizes  $r_0 - r_{\text{split}}$ . Before proceeding with the derivation, we introduce several definitions and state a few properties that will be useful later in the derivation. Let  $\phi(s)$  denote the standard Gaussian probability density function, and let  $\Phi(s)$  denote the cumulative distribution function of a standard Gaussian. It is easy to show that

$$\phi'(s) = -s\phi(s), \quad (18)$$

$$\Phi'(s) = \phi(s). \quad (19)$$

Our objective function  $\max_t r_0 - r_{\text{split}}$  can be written as

$$\max_t \frac{n_1 n_2}{n} \|\bar{u}_1 - \bar{u}_2\|^2 = n \left( \max_t \frac{n_1 n_2}{n^2} \|\bar{u}_1 - \bar{u}_2\|^2 \right).$$

Assuming a large number of measurements,

$$\begin{aligned} \frac{n_1}{n} &\approx \int_{-\infty}^t \phi(s) ds = \Phi(t), \\ \frac{n_2}{n} &\approx 1 - \Phi(t), \end{aligned}$$

$$\begin{aligned}\bar{u}_1 &\approx \mathbb{E}[u \mid s < t] = \frac{\int_{-\infty}^t s\phi(s) ds}{\int_{-\infty}^t \phi(s) ds} \stackrel{(18)(19)}{=} \frac{-\phi(t)}{\Phi(t)} v, \\ \bar{u}_2 &\approx \mathbb{E}[u \mid s \geq t] = \frac{\int_t^{\infty} s\phi(s) ds}{\int_t^{\infty} \phi(s) ds} \stackrel{(18)(19)}{=} \frac{\phi(t)}{1 - \Phi(t)} v.\end{aligned}$$

We substitute the above equalities into the objective function:

$$\begin{aligned}\max_t \frac{n_1 n_2}{n^2} \|\bar{u}_1 - \bar{u}_2\|^2 \\ \approx \max_t \Phi(t) [1 - \Phi(t)] \phi^2(t) \left[ -\frac{1}{\Phi(t)} - \frac{1}{1 - \Phi(t)} \right]^2 \|v\|^2 \\ = \max_t \frac{\phi^2(t)}{\Phi(t) [1 - \Phi(t)]} = \max_t h(t).\end{aligned}\tag{20}$$

Using the following properties:

$$\phi(-t) = \phi(t), \tag{21}$$

$$\Phi(-t) = 1 - \Phi(t), \tag{22}$$

$h(t)$  is symmetric about  $t = 0$ . To show that  $\arg \max_t h(t) = 0$ , it is sufficient to show (i)  $\left. \frac{d \log h(t)}{dt} \right|_{t=0} = 0$ , and (ii)

$\left. \frac{d \log h(t)}{dt} \right|_{t>0} < 0$ . Taking the derivative:

$$\begin{aligned}\log h(t) &\stackrel{(22)}{=} 2 \log \phi(t) - \log \Phi(t) - \log \Phi(-t), \\ \frac{d \log h(t)}{dt} &\stackrel{(18)(19)}{=} -2 \frac{t\phi(t)}{\phi(t)} - \frac{\phi(t)}{\Phi(t)} + \frac{\phi(t)}{\Phi(-t)} \\ &= -2t + \phi(t) \left[ \frac{1}{\Phi(-t)} - \frac{1}{\Phi(t)} \right].\end{aligned}\tag{23}$$

From (23), it is easy to check that  $\left. \frac{d \log h(t)}{dt} \right|_{t=0} = 0$ . Next, we show  $\left. \frac{d \log h(t)}{dt} \right|_{t>0} < 0$ , i.e.,

$$\phi(t) \left[ \frac{1}{\Phi(-t)} - \frac{1}{\Phi(t)} \right] < 2t \quad \text{for } t > 0. \tag{24}$$

We define the inverse Mills ratio,

$$\lambda(t) = \frac{\phi(t)}{\Phi(-t)}, \quad \lambda(-t) \stackrel{(21)}{=} \frac{\phi(t)}{\Phi(t)}.$$

Therefore,

$$\begin{aligned}\phi(t) \left[ \frac{1}{\Phi(-t)} - \frac{1}{\Phi(t)} \right] &= \lambda(t) - \lambda(-t) \\ &= \int_{-t}^t \lambda'(x) dx \quad \text{for } t > 0.\end{aligned}\tag{25}$$

From (25), to show that  $\left. \frac{d \log h(t)}{dt} \right|_{t>0} < 0$ , or equivalently (24), we require  $\lambda'(t) < 1$  and  $t > 0$  (so that the integral from  $-t$  to  $t$  is valid). We therefore take the derivative of  $\lambda(t)$ :

$$\begin{aligned}\lambda'(t) &= \frac{\phi'(t)\Phi(-t) + \phi(t)\Phi'(-t)}{\Phi(-t)^2} \\ &\stackrel{(18)(19)}{=} \frac{-t\phi(t)\Phi(-t) + \phi(t)\phi(t)}{\Phi(-t)^2} \\ &= \frac{\phi(t)}{\Phi(-t)} \left[ \frac{\phi(t)}{\Phi(-t)} - t \right] \\ &= \lambda(t) [\lambda(t) - t].\end{aligned}\tag{26}$$

Note the following equalities:

$$\mathbb{E}[s \mid s \geq t] = \frac{\int_t^{\infty} s\phi(s) ds}{\int_t^{\infty} \phi(s) ds} \stackrel{(18)(19)(22)}{=} \frac{\phi(t)}{\Phi(-t)} = \lambda(t).$$

$$\begin{aligned}\mathbb{E}[s^2 \mid s \geq t] &= \frac{\int_t^{\infty} s^2 \phi(s) ds}{\int_t^{\infty} \phi(s) ds} = \frac{\int_t^{\infty} s^2 \phi(s) ds}{\Phi(-t)} \\ &= \frac{-s\phi(s)|_t^{\infty} + \int_t^{\infty} \phi(s) ds}{\Phi(-t)} \\ &= \frac{t\phi(t) + \Phi(-t)}{\Phi(-t)} \\ &= t\lambda(t) + 1,\end{aligned}$$

where we have used integration by parts. Further,

$$\begin{aligned}\text{Var}(s \mid s \geq t) &= \mathbb{E}[s^2 \mid s \geq t] - \mathbb{E}[s \mid s \geq t]^2 \\ &= t\lambda(t) + 1 - \lambda(t)^2 > 0,\end{aligned}$$

since  $\text{Var}(s \mid s \geq t) > 0$ . The above inequality leads to

$$\lambda(t)^2 - t\lambda(t) \stackrel{(26)}{=} \lambda'(t) < 1.$$

Therefore, using (25),

$$\begin{aligned}\phi(t) \left[ \frac{1}{\Phi(-t)} - \frac{1}{\Phi(t)} \right] &= \lambda(t) - \lambda(-t) \\ &= \int_{-t}^t \lambda'(x) dx < \int_{-t}^t 1 = 2t \quad \text{for } t > 0.\end{aligned}$$

Consequently, (24) holds and  $\left. \frac{d \log h(t)}{dt} \right|_{t>0} < 0$  holds. We conclude that  $\arg \max_t h(t) = 0$ .

Next, we show that the half-space splits are optimal among all splits. We continue to assume a large number of measurements and generalize the previous derivation to arbitrary splits:

$$\begin{aligned}\frac{n_1}{n} &\approx \int_{G_1} \phi(u) d^d u = \mathbb{P}(G_1) = p, \\ \frac{n_2}{n} &\approx 1 - p,\end{aligned}$$

$$\begin{aligned}\bar{u}_1 &\approx \mathbb{E}[u \mid u \in G_1] = \frac{\int_{G_1} u\phi(u) d^d u}{\int_{G_1} \phi(u) d^d u} = \frac{\int_{G_1} u\phi(u) d^d u}{p}, \\ \bar{u}_2 &\approx \mathbb{E}[u \mid u \in G_2] = \frac{\int_{G_2} u\phi(u) d^d u}{\int_{G_2} \phi(u) d^d u} = \frac{\int_{G_2} u\phi(u) d^d u}{1 - p}.\end{aligned}$$

By the law of total expectation,

$$\begin{aligned}p\mathbb{E}[u \mid u \in G_1] + (1 - p)\mathbb{E}[u \mid u \in G_2] \\ = \int_{\mathbb{R}^d} u\phi(u) d^d u = \mathbb{E}[u] = 0.\end{aligned}$$

Using the equalities above, the objective function can be written as

$$\begin{aligned}
& \max_{G_1} \frac{n_1 n_2}{n^2} \|\bar{u}_1 - \bar{u}_2\|^2 \\
& \approx \max_{G_1} p(1-p) \left\| \mathbb{E}[u \mid u \in G_1] - \frac{p}{p-1} \mathbb{E}[u \mid u \in G_1] \right\|^2 \\
& = \max_{G_1} \frac{p}{1-p} \|\mathbb{E}[u \mid u \in G_1]\|^2 \\
& = \max_p \max_{G_1: \mathbb{P}(G_1)=p} \frac{p}{1-p} \|\mathbb{E}[u \mid u \in G_1]\|^2,
\end{aligned}$$

where the last equality decomposes the maximization into two steps and makes the following observation clearer: for a fixed  $p$ , we need to maximize  $\|\mathbb{E}[u \mid u \in G_1]\|$  in order to maximize the objective function. Let  $v$  be:

$$v = \frac{\mathbb{E}[u \mid u \in G_1]}{\|\mathbb{E}[u \mid u \in G_1]\|}. \quad (27)$$

Substituting in  $v$  yields:

$$\begin{aligned}
& \max_{G_1: \mathbb{P}(G_1)=p} p \|\mathbb{E}[u \mid u \in G_1]\| \\
& = \max_{G_1: \mathbb{P}(G_1)=p} p (v^\top \mathbb{E}[u \mid u \in G_1]) \\
& = \max_{G_1: \mathbb{P}(G_1)=p} \int_{G_1} v^\top u \phi(u) \, d^d u \\
& \stackrel{s=v^\top u}{=} \max_{G_1: \mathbb{P}(G_1)=p} \int_{G_1} s \phi(u) \, d^d u \\
& = \max_{G_1: \mathbb{P}(G_1)=p} \int_{s: (s,z) \in G_1} \left[ \int_{z: (s,z) \in G_1} \phi(z) \, d^{d-1} z \right] s \phi(s) \, ds \\
& = \max_{G_1: \mathbb{P}(G_1)=p} \int_{s: (s,z) \in G_1} s \left[ \int_{\mathbb{R}^{d-1}} \phi(z) \, d^{d-1} z \right] \phi(s) \, ds,
\end{aligned}$$

where the penultimate equality decomposes the integral into the coordinate along  $v$  and the remaining  $d-1$  orthogonal coordinates, using the isotropic Gaussian assumption for  $u$  and the resulting independence between coordinates. The last equality considers the following maximization process: starting from the largest possible value of  $s$ , we include all points in the orthogonal coordinates to assign as much probability mass as possible to each value of  $s$  before including points with smaller values of  $s$ . The process continues until the total probability mass equals  $p$ . Therefore,

$$\begin{aligned}
& \arg \max_{G_1: \mathbb{P}(G_1)=p} \|\mathbb{E}[u \mid u \in G_1]\| \\
& = \arg \max_{G_1: \mathbb{P}(G_1)=p} \int_{\{s \mid \forall z \in \mathbb{R}^{d-1}, (s,z) \in G_1\}} s \phi(s) \, ds \quad (28) \\
& = \{u : v^\top u = s \geq t\}, \quad (29)
\end{aligned}$$

where  $t$  is selected such that  $\mathbb{P}(G_1) = p$ . Note that (28) is essentially a one-dimensional problem, so it is obvious that we need to include the largest possible values of  $s$  to maximize the objective, leading to  $s \geq t$  as the solution. It is also worthwhile to mention that  $v$  is defined in (27) as a way to rewrite  $\|\mathbb{E}[u \mid u \in G_1]\|$  for the derivation. Because of the isotropic Gaussian assumption,  $v$  may point in any direction. In other words, the solution is not unique. Any half-space split satisfying (29) and  $\mathbb{P}(G_1) = p$  is optimal.