Robust Incremental Online Inference Over Sparse Factor Graphs: Beyond the Gaussian Case

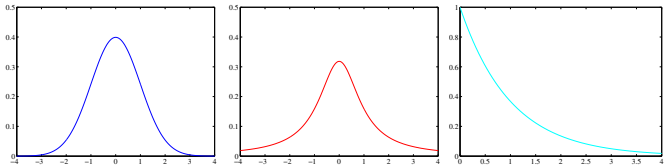
David M. Rosen, Michael Kaess, and John J. Leonard

***Abstract*—Many online inference problems in robotics and AI are characterized by probability distributions whose factor graph representations are sparse. While there do exist some computationally efficient algorithms (e.g. incremental smoothing and mapping (iSAM) or Robust Incremental least-Squares Estimation (RISE)) for performing online incremental maximum likelihood estimation over these models, they generally require that the distribution of interest factors as a product of Gaussians, a rather restrictive assumption. In this paper, we investigate the possibility of performing efficient incremental online estimation over sparse factor graphs in the non-Gaussian case. Our main result is a method that generalizes iSAM and RISE by removing the assumption of Gaussian factors, thereby significantly expanding the class of distributions to which these algorithms can be applied. The generalization is achieved by means of a simple algebraic reduction that under relatively mild conditions (boundedness of each of the factors in the distribution of interest) enables an instance of the general maximum likelihood estimation problem to be reduced to an equivalent instance of least-squares minimization that can be solved efficiently online by application of iSAM or RISE. Through this construction we obtain robust, computationally efficient, and *mathematically correct* incremental online maximum likelihood estimators for non-Gaussian distributions over sparse factor graphs.**

I. Introduction

Many online inference problems in robotics and AI are characterized by probability distributions whose factor graph representations are sparse; for example, both bundle adjustment [6], [25] and the smoothing formulation of simultaneous localization and mapping (SLAM) [23], [24] belong to this class. Online maximum likelihood estimation over these models corresponds to solving a sequence of maximization problems in which the objective function is a product to which additional factors are appended over time. In practice, these problems are often solved by computing each estimate in the sequence as the solution of an independent maximization problem using standard iterative numerical techniques. While this approach is general and produces good results, it is computationally expensive, and does not exploit the sequential nature of the underlying inference problem; this limits its utility in real-time online applications, where speed is critical.

In previous work [20] Rosen *et al.* developed Robust Incremental least-Squares Estimation (RISE), an improved version of incremental smoothing and mapping (iSAM [9], [10]) obtained by incrementalizing the Powell's Dog-Leg trust-region method [18], to efficiently solve the online maximum likelihood estimation problem for the special case in which the distribution of interest factors as a product of Gaussians. In this special case, maximum likelihood estimation is equivalent to minimizing a sum of squared measurement errors. RISE and its predecessor iSAM achieve efficient computation in this case by exploiting structural properties of sequential least-squares minimization (namely, an efficient mechanism for updating a sparse QR decomposition of the Jacobian of the error residual function, cf. [20, Sec. 3]) to produce a *direct update* to the previously computed estimate when new data arrives, rather than recomputing a new estimate from scratch. This incremental approach enables iSAM and RISE to achieve computational speeds unmatched by iterated batch techniques (such as the popular Levenberg-Marquardt) while producing solutions of comparable accuracy.



(a) Gaussian (b) Cauchy (c) Exponential

Fig. 1. Examples of common probability density functions. While the assumption of additive Gaussian noise is ubiquitous in robotics and computer vision applications, in reality system noise often follows a non-Gaussian distribution. Gaussian models are particularly ill-suited for approximating heavy-tailed distributions like the Cauchy distribution or asymmetric, one-sided distributions like the exponential. The unprincipled use of Gaussian models in the presence of such perturbations can lead to poor system performance or outright failure. Achieving robust long-term autonomy requires the development of computationally efficient online inference methods that are able to correctly address non-Gaussian error distributions.

However, while the assumption of additive Gaussian noise is ubiquitous in robotics and computer vision due to its computational expedience, in reality there are many distributions that arise in practice (e.g. the exponential [2], [21] or Cauchy [15] distributions) that are not at all well-modeled as Gaussians (Fig. 1). Multi-modal, asymmetric, or one-sided distributions like the exponential are difficult to capture with a unimodal, symmetric, and everywhere-nonzero Gaussian model. Heavy-tailed distributions like the Cauchy distribution are particularly pernicious; while the Cauchy and Gaussian distributions may appear visually quite similar, their tail behaviors differ so radically that performing estimation in the presence of Cauchy-distributed noise using a Gaussian model is essentially hopeless. Clearly, achieving practical robust long-term autonomy in the presence of these kinds of perturbations requires the development of efficient online inference methods that do not depend upon the Gaussian assumption.

The authors are with the Massachusetts Institute of Technology, Cambridge, MA 02139, USA {dmrosen,kaess,jleonard}@mit.edu

Unfortunately, the incrementalization procedure used to obtain iSAM and RISE from the Gauss-Newton and Powell's Dog-Leg optimization methods (respectively) depends crucially upon the appearance of the Jacobian in these algorithms, which are restricted to sum-of-squares objective functions; it does not extend to the more general Newton-type optimization methods, where the Jacobian is replaced by the Hessian.

To that end, in this paper we consider an alternative approach for generalizing iSAM and RISE beyond the Gaussian case: rather than generalizing the optimization methods themselves beyond quadratic cost functions, we instead show how to reduce an instance of the general maximum likelihood estimation problem to an equivalent instance of least-squares minimization, which can then in turn be solved efficiently online by direct application of iSAM or RISE. Through this construction we obtain robust, computationally efficient, and *mathematically correct* incremental online maximum likelihood estimators for non-Gaussian distributions over sparse factor graphs.

II. Generalizing iSAM/RISE

In this section we generalize iSAM and RISE by removing the assumption of Gaussian factors. This generalization is achieved by means of Theorem 1, which provides a method for reducing an instance of maximization of a product of functions to an equivalent instance of least-squares minimization that can be solved efficiently online by application of iSAM or RISE.

A. Maximum likelihood estimation in factor graphs

A *factor graph* is a bipartite graph $G = (\mathcal{F}, \Theta, \mathcal{E})$ with two node types: *factor nodes* $f_i \in \mathcal{F}$ (each representing a real-valued function) and *variable nodes* $\theta_j \in \Theta$ (each representing an argument to one or more of the functions in $\mathcal{F}$). Every real-valued function $f\colon \Omega \to \mathbb{R}$ has a corresponding factor graph G encoding its factorization as

$$\begin{aligned} f(\Theta) &= \prod_{i=1}^{m} f_i(\Theta_i), \\ \Theta_i &= \{\theta_j \in \Theta \mid (f_i, \theta_j) \in \mathcal{E}\}. \end{aligned} \tag{1}$$

If a function $p\colon \Omega \to \mathbb{R}$ that factors as in equation (1) is a probability distribution, then maximum likelihood estimation corresponds to finding the value Θ^* that maximizes (1):

$$\Theta^* = \operatorname*{argmax}_{\Theta \in \Omega} p(\Theta) = \operatorname*{argmax}_{\Theta \in \Omega} \prod_{i=1}^{m} p_i(\Theta_i). \tag{2}$$

B. Reducing maximization of a product of functions to an equivalent instance of least-squares minimization

The extension of iSAM and RISE to the problem of general maximum likelihood estimation (2) is achieved by means of the following theorem, which shows that under relatively mild conditions (namely, boundedness and positivity of the factors), maximization of a product of functions is equivalent to an instance of least-squares minimization.

Theorem 1 (Reducing MLE to least-squares minimization): Let $f\colon \Omega \to \mathbb{R}$ be a function that factors as

$$f(x) = \prod_{i=1}^{m} f_i(x), \tag{3}$$

and suppose that $f_i > 0$ and $\|f_i\|_\infty < \infty$ for all $1 \le i \le m$. Fix constants $c_i \in \mathbb{R}$ such that $c_i > \|f_i\|_\infty$ and define

$$\begin{aligned} r_i &\colon \Omega \to \mathbb{R} \\ r_i(x) &= \sqrt{\ln c_i - \ln f_i(x)} \end{aligned} \tag{4}$$

for all $1 \le i \le m$. Then

$$\operatorname*{argmax}_{x \in \Omega} f(x) = \operatorname*{argmin}_{x \in \Omega} \sum_{i=1}^{m} r_i(x)^2. \tag{5}$$

Proof: Define

$$\tilde{f}_i(x) = \frac{1}{c_i} \cdot f_i(x).$$

Then $\tilde{f}_i > 0$ and $\|\tilde{f}_i\|_\infty < 1$ for all $1 \le i \le m$, and

$$\begin{aligned} \operatorname*{argmax}_{x \in \Omega} f(x) &= \operatorname*{argmax}_{x \in \Omega} \prod_{i=1}^{m} f_i(x) \\ &= \operatorname*{argmax}_{x \in \Omega} \prod_{i=1}^{m} \tilde{f}_i(x) \\ &= \operatorname*{argmax}_{x \in \Omega} \sum_{i=1}^{m} \ln \tilde{f}_i(x) \\ &= \operatorname*{argmin}_{x \in \Omega} \sum_{i=1}^{m} -\ln \tilde{f}_i(x). \end{aligned} \tag{6}$$

Since $0 < \tilde{f}_i(x_i) < 1$, then $-\ln \tilde{f}_i(x) > 0$, and therefore the function $r_i\colon \Omega \to \mathbb{R}$ specified by

$$r_i(x) = \sqrt{-\ln \tilde{f}_i(x)} = \sqrt{\ln c_i - \ln f_i(x)} \tag{7}$$

is well-defined for all $1 \le i \le m$. Substitution of (7) into the final line of (6) proves the result. ∎

C. Application to online maximum likelihood estimation over sparse factor graphs

In this subsection we show how to apply Theorem 1 to the probability distribution p in (2).

So far, everything that we have said has been perfectly general; henceforward, we will assume that $\Omega \subseteq \mathbb{R}^n$, that p is a probability density function on Ω, and that $p_i \in C^1(\Omega)$ (so that we can apply gradient-based optimizers such as iSAM and RISE) and satisfies the hypotheses of Theorem 1 for all $1 \le i \le m$ (we will discuss how restrictive these assumptions are as a practical matter in Section III). Under these conditions, we may apply the reduction given in Theorem 1 to obtain the equivalent maximum likelihood estimate Θ^* as

$$\Theta^* = \operatorname*{argmin}_{\Theta \in \Omega} \sum_{i=1}^{m} r_i(\Theta_i)^2 = \operatorname*{argmin}_{\Theta \in \Omega} \|r(x)\|^2 \tag{8}$$

for $r\colon \Omega \to \mathbb{R}^m$.

Note that each of the factors p_i of p in (2) gives rise to a corresponding summand r_i in the least-squares minimization (8) by means of equation (4); consequently, performing *online* inference, in which new measurements become available (equivalently, in which new factor nodes are added to G) over time, corresponds to solving a *sequence* of minimization problems of the form (8) in which new summands are added to the objective function over time. If G is sparse, then the corresponding sequential least-squares minimizations (8) are likewise sparse. By this construction, we thus reduce the general problem of sequential maximum likelihood estimation over sparse factor graphs (2) to an equivalent instance of sequential sparse least-squares minimization (8), which can be solved efficiently online by application of iSAM or RISE.

III. How restrictive are the hypotheses of Theorem 1?

The method of generalizing iSAM/RISE developed in Section II depends upon the application of Therorem 1; consequently, this generalization only extends as far as does Theorem 1 itself. In this section, we consider the restrictiveness of the hypotheses of Theorem 1, and show that (at least as a practical matter) they in fact hold with great generality.

To see this, suppose that the function p defined by (2) is a probability distribution. Since $p \geq 0$, the positivity hypothesis of Theorem 1 can always be satisfied, if necessary by restricting attention to the open set $\mathring{\Omega} = \{x \in \Omega \mid p(x) > 0\}$ (which contains all of the physically relevant points in the distribution's domain). If each of the factors p_i additionally happens to be bounded on Ω (which will always be the case if Ω is compact), then the reduction proceeds.

Furthermore, we also claim that *even in cases where some of the factors p_i may be unbounded*, it will often be possible to produce a reasonable approximation to which Theorem 1 *does* apply. This approximation is straightforward: it simply consists of the recognition that, for many problems of interest in science and engineering, there is a (perhaps very large) bounded set of physically plausible values within which the parameters to be estimated ought to fall. For example, when estimating the location of a robotic ground vehicle, it is entirely reasonable to restrict attention to the surface of the Earth, even though the formal probability density function describing the vehicle's location may have nonzero (albeit very very small) amplitude in the vicinity of, say, Saturn.

More formally, if it is possible to use *a priori* knowledge of the problem at hand to restrict attention to a (possibly very large) compact set $C \subseteq \Omega$, then the fact that $p_i \in C^1(\Omega)$ implies that p_i is bounded on C for all $1 \leq i \leq m$, and therefore Theorem 1 applies to $p|_C$. Furthermore, if the probability distribution that we wish to maximize happens to be the posterior distribution for a parameter with known prior p_0 (as in Bayesian inference), then given any $\epsilon > 0$, we can always find a compact set $C \subseteq \Omega$ such that $1 - p_0(C) < \epsilon$, i.e., such that the prior probability that the true parameter value lies outside of C (hence will be missed by estimation over the restricted function $p|_C$) will be less than ϵ. In other words, this method admits the direct control of the tightness of the approximation (as measured by ϵ) through the selection of the compact set C.

Finally, we also observe that although Theorem 1 requires the identification of an upper bound c_i for each factor p_i, this bound need not be *tight* (although tighter choices of c_i do give rise to better numerical properties in the algorithm). This is also quite convenient, as it is often the case in practice that it is difficult or impossible to provide a closed-form analytic solution for the absolute maximum of a given function, but relatively easy to produce at least a coarse upper bound.

Thus, we see that the hypotheses of Theorem 1 are sufficiently mild (when coupled with some physically motivated yet still principled approximation, if necessary) to admit its application to a wide variety of problems of interest.

IV. Experimental Results

In this section we illustrate the utility of the newly-generalized RISE algorithm by comparing its performance with that of its predecessor (Gaussian-only RISE) on two example applications inspired by marine robotics.

A. Example application: Measuring acoustic time-of-flight

We first consider a relatively straightforward estimation problem from marine robotics in order to highlight the insufficiency of the Gaussian assumption in the general case.

The underwater environment is extremely acoustically noisy; so much so, in fact, that it has been hypothesized that time-of-flight measurement errors in long-baseline (LBL) underwater localization systems may follow a Cauchy distribution [15]. Suppose that this is so, and consider a perfectly stationary autonomous underwater vehicle (AUV) attempting to measure the acoustic time-of-flight to a particular sonar beacon through repeated interrogation.

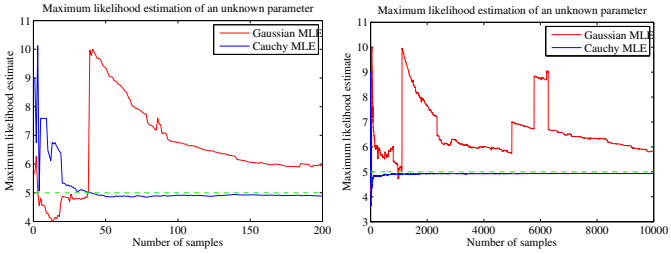
Mathematically, this corresponds to the simplest possible inference problem: estimating the value of an unknown parameter a given a sequence $Y_N = (y_1, \ldots, y_N)$ of measurements corrupted by additive i.i.d. noise. We suppose that this noise is generated by a Cauchy distribution, so that

$$y_i = a + e_i, \quad e_i \sim C(x_0, \gamma) \quad \forall 1 \leq i \leq N \tag{9}$$

where here we use the notation $X \sim C(x_0, \gamma)$ to indicate that the random variable X follows a Cauchy distribution with location parameter x_0 and shape parameter γ, corresponding to the probability density function

$$p_X(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}. \tag{10}$$

Now if we are restricted to using only Gaussian models for inference (as we are in the case of vanilla iSAM or RISE), then the best that we can do in this case is to model the noise process (9) as x_0-mean Gaussian. Given the sequence Y_N, the



(a) The first 200 samples (b) The full 10000 samples

Fig. 2. Estimation of an unknown parameter: This plot shows the maximum likelihood estimates for the parameter a as a function of the number of observations y_i sampled from (9) for $x_0 = 0$, $\gamma = 1$, and $a = 5$. The red line corresponds to the estimator $\hat{a}^G_{ML}$ in (11) obtained under the Gaussian assumption, while the blue line corresponds to the incrementalized estimator $\hat{a}^C_{ML}$ obtained from equation (13) using the method of Section II-C and run online using RISE2. The dashed green line shows the true value of the parameter a.

maximum likelihood estimator $\hat{a}^G_{ML}$ for a under this Gaussian assumption is then just the x_0-translated sample mean:

$$\hat{a}^G_{ML}(Y_N) = \frac{1}{N}\sum_{i=1}^{N} y_i - x_0. \tag{11}$$

Let us consider the performance of the estimator (11). We observe that since each of the y_i are random variables, then so is $\hat{a}^G_{ML}$ itself; furthermore, since we know the distribution (9) from which the y_i are sampled, we can compute the distribution for $\hat{a}^G_{ML}$. Using the fact that $\hat{a}^G_{ML}(Y_N)$ is a linear combination of the i.i.d. random variables y_i [17, Secs. 4.4 and 5.4] together with a little algebra [1] shows that

$$\hat{a}^G_{ML} \sim C(a, \gamma). \tag{12}$$

Equation (12) shows that the distribution for the estimator $\hat{a}^G_{ML}$ for a derived under the Gaussian assumption is *independent* of the number of observations y_i collected. In other words, *no matter how much data is accumulated, the uncertainty in the estimate $\hat{a}^G_{ML}$ will never decrease* (in particular, the estimator $\hat{a}^G_{ML}$ is *inconsistent*).

On the other hand, consider the maximum likelihood estimator $\hat{a}^C_{ML}$ for a obtained using the correct (Cauchy) noise model from (9):

$$\hat{a}^C_{ML}(Y_N) = \operatorname*{argmax}_{x\in\mathbb{R}} \frac{1}{\gamma^N \pi^N} \prod_{i=1}^{N} \left[1 + \left(\frac{y_i - x - x_0}{\gamma}\right)^2\right]^{-1}. \tag{13}$$

In contrast to $\hat{a}^G_{ML}$, this estimator is consistent, asymptotically normal, and its variance scales asymptotically as $1/N$ [3] (as we would hope, since this is the mathematically correct maximum likelihood estimator).

Now we compare the performance of these two estimators in practice. For this experiment we sampled 10000 observations from the distribution (9) with $x_0 = 0$, $\gamma = 1$, and $a = 5$. This data set was processed twice: once using the estimator $\hat{a}^G_{ML}$ defined in (11), and again using the incrementalized estimator $\hat{a}^C_{ML}$ obtained from equation (13) using the method of Section II-C and run online using the RISE2 [20] implementation in Georgia Tech's GTSAM library (version 2.0.0, available through https://collab.cc.gatech.edu/borg/gtsam/) with the default settings. The resulting estimates are shown in Fig. 2 as a function of the number of samples observed. We can clearly see that while the correct estimator $\hat{a}^C_{ML}$ implemented using the method of Section II-C converges rapidly to the true value as expected (Fig. 2(a)), the estimator $\hat{a}^G_{ML}$ obtained under the Gaussian assumption shows no convergence towards the true value whatsoever, as predicted by (12) (Fig. 2(b)).

The failure of the Gaussian estimator $\hat{a}^G_{ML}$ to produce a good estimate for a is due to the fact that the Cauchy distribution (10) is *fat-tailed*; that is, its density p_X satisfies

$$\lim_{x\to\pm\infty} \left(|x|^{\alpha+1} \cdot p_X(x)\right) = 1 \text{ for some } 0 < \alpha < 2.$$

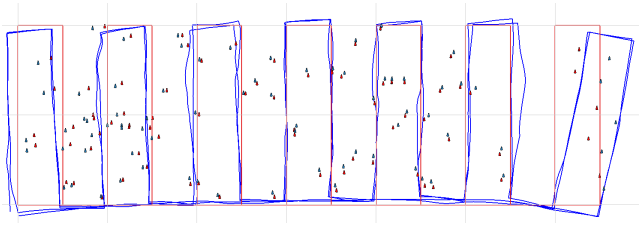
Distributions of this type do not have well-defined variances (intuitively, they have infinite variances), and therefore their sample averages do *not* satisfy the hypotheses of the classical central limit theorem (that is, the distribution of their sample averages is *not* asymptotically normal). There is a generalization of the classical central limit theorem due to Gnedenko and Kolmogorov [5] which implies that sample averages of i.i.d. variables drawn from a fat-tailed distributon with a given α parameter will converge to a member of the Lévy α-stable family having the same α parameter; this explains why $\hat{a}^G_{ML}$ is Cauchy-distributed, as the Cauchy distributions are members of the α-stable family with $\alpha = 1$. However, fat-tailed distributions do not have well-defined means for $\alpha \in (0, 1]$. Thus, while the stability property (12) of the estimator $\hat{a}^G_{ML}$ is a special consequence of the fact that the observations in this case are drawn from a Cauchy distribution, the failure of convergence of the Gaussian estimator $\hat{a}^G_{ML}$ is *generic*: this failure will occur whenever observations are sampled from a fat-tailed distribution with $\alpha \leq 1$ simply because, in those circumstances, the underlying distribution has no mean for the sample average to converge *to*. Even assuming that $\alpha > 1$, in which case the estimator $\hat{a}^G_{ML}$ has a well-defined expectation, its variance will still be infinite for $\alpha < 2$, rendering it unsuitable for use as a practical matter.

In contrast, mathematically correct maximum likelihood estimators are consistent (i.e. will converge in probability to the correct result as $N \to \infty$) and asymptotically normal *in the general case* under reasonably mild conditions, with (co)variance scaling asymptotically as $1/N$ (cf. Theorems 17 and 18 in [4] for details).

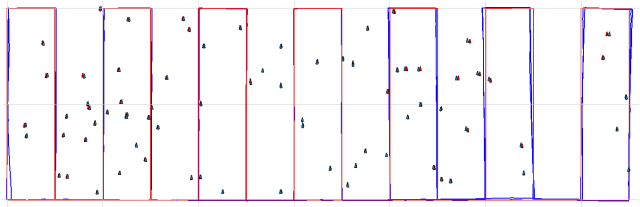
B. Example application: Underwater SLAM

The simplicity of the example application in Section IV-A was designed to highlight the mathematical underpinnings of the insufficiency of the Gaussian assumption in the general case. In this section, we consider a more realistic example of how the method of Section II-C would be applied in practice.

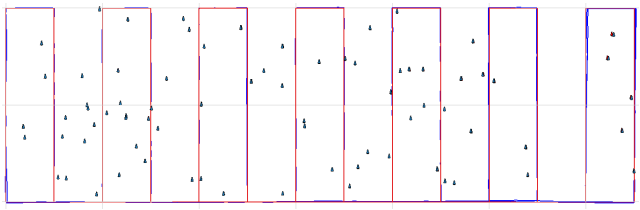
For this experiment we consider an AUV searching the bottom of a channel for targets of interest. We assume that the AUV is equipped with a forward-looking sonar (with a maximum effective range of 40 meters and a horizontal field of view of 90 degrees) and an inertial measurement unit (IMU)



(a) Nominal Gaussian model p_s



(b) Optimal Gaussian model p_g



(c) Mixture model p_r

Fig. 3. Estimated AUV trajectory and sonar target locations (blue) versus ground truth (red) for each of the sonar ranging error density estimates. (a) Estimate obtained using the nominal Gaussian model p_s (median sonar target positioning error: 3.29 m). (b) Estimate obtained using the optimal Gaussian model p_g (median sonar target positioning error: .453 m). (c) Estimate obtained using the mixture model p_r (median sonar target positioning error: .126 m). Note the superior sharpness/fine detail of (c) versus (b). The grid spacing is 50 m.

and Doppler velocity log (DVL) for ego-motion estimation. We further suppose that there are 75 sonar targets uniformly randomly distributed over the AUV's area of operation, and that the AUV searches for these by following a standard lawnmower search pattern at a depth of 7 meters (cf. Fig. 3). The AUV emits sonar pings at regular intervals (once for every 2.5 meters traveled) and simultaneously produces an estimate of its ego-motion (translation and rotation) since the previous ping by integrating its IMU and DVL measurements.

We simulate sensor errors as follows. Errors in the AUV's ego-motion estimate are Gaussian and mean-zero, with a standard deviation of .02 m along each translational axis and 1 degree standard deviation in heading. The error in the sonar receiver is also modeled as mean-zero Gaussian noise, with a standard deviation of $\sigma_r = .1$ m in range and 1 degree in bearing. We also simulate errors in the detection process: every time the AUV emits a ping, any target within the sonar's field of view is only detected with probability .8, and each detection originates from a multipath reflection off the water's surface with probability $M = .25$.

Let us consider the probability density function p_r for the error in the sonar's range measurement. According to the above description, p_r is formally given by:

$$p_r(x) = \underbrace{(1-M)\cdot p_s(x)}_{\text{internal receiver ranging error}} + \underbrace{M\cdot(p_s * p_m)(x)}_{\text{sum of internal and multipath errors}}, \tag{14}$$

where

$$p_s(x) = \frac{1}{\sqrt{2\pi}\sigma_r}\exp\left(-\frac{x^2}{2\sigma_r^2}\right) \tag{15}$$

is the density for the ranging error due to the sonar receiver itself, and p_m is the probability density function describing the distribution of range errors due to multipath reflection. While we do not know the explicit closed-form pdf for p_m, we do at least know that it is supported on $[0,\infty)$ (since the multipath path length always exceeds the true range to target), and we may draw samples from this distribution by running our previously described simulator. Therefore, we appeal to the principle of maximum entropy [8], [22], and model p_m as the maximum entropy distribution $\hat{p}_m$ supported on $[0,\infty)$ with mean and variance equal to the sample mean μ_m and sample variance σ_m^2 for a set of multipath reflection errors sampled from the sonar simulator. Standard techniques from the calculus of variations then show that $\hat{p}_m$ has the parametric form:

$$\hat{p}_m(x) = \begin{cases} c\exp\left(-\lambda_1 x - \lambda_2 x^2\right), & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{16}$$

where the parameters c, λ_1, and λ_2 are chosen so that $\hat{p}_m$ is properly normalized and has mean μ_m and variance σ_m^2.

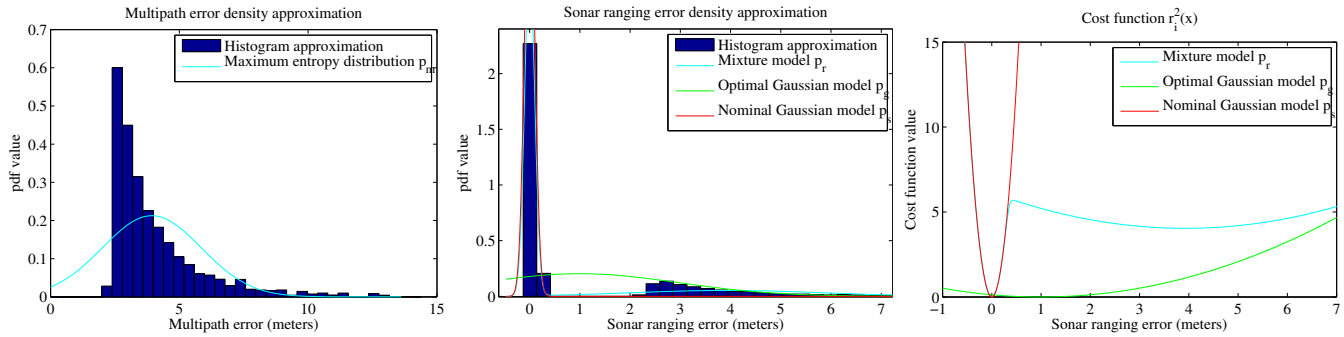
Given (15) and (16), we may compute the closed form for the convolution in (14):

$$(p_s * \hat{p}_m)(x) = \frac{c}{2\sqrt{1+2\lambda_2\sigma_r^2}}\left(1+\operatorname{erf}\left(\frac{x-\lambda_1\sigma_r^2}{\sigma_r\sqrt{2+4\lambda_2\sigma_r^2}}\right)\right) \cdot \exp\left(\frac{-2\lambda_2 x^2 - 2\lambda_1 x + \lambda_1^2\sigma_r^2}{2+4\lambda_2\sigma_r^2}\right). \tag{17}$$

Equations (15) and (17) together completely specify the pdf p_r that we will use to model the sonar's ranging error.

Now obviously p_r is non-Gaussian. But what if we are restricted to use only Gaussian models, as in the case of vanilla iSAM/RISE? One common approach is simply to use the nominal sonar model p_s in (15), completely disregarding the possibility of multipath errors. A better model p_g (indeed, the information-theoretically optimal Gaussian model) is obtained by moment-matching the mean and variance of a sample of ranging errors drawn from the complete (i.e. internal receiver and multipath) sonar error distribution.

To determine the models used in the experiment, 4000 multipath reflection errors were drawn from the sonar error simulator described above, and the sample mean $\mu_m = 4.03$ and variance $\sigma_m^2 = 3.27$ determined. The parameter values for the maximum entropy distribution $\hat{p}_m$ in (16) with these moments were then determined numerically in MATLAB, producing $c = 2.59\times10^{-2}$, $\lambda_1 = -1.07$, $\lambda_2 = .136$; this data determines the complete mixture model p_r in (14)



(a) Multipath error density approximation (b) Sonar ranging error density approximations (c) Sonar ranging error cost functions

Fig. 4. Sonar ranging error models used in this experiment. (a) The maximum entropy density $\hat{p}_m$ used to estimate the multipath ranging error density (obtained by sampling 4000 multipath ranging errors from the sonar error simulator) together with the histogram density estimate. (b) The three models used to estimate the complete (internal and multipath) sonar ranging error density: the mixture model p_r (cyan, determined by the maximum entropy density $\hat{p}_m$ shown in (a)), the optimal Gaussian model p_g (green, obtained by moment-matching a sample of 16000 sonar ranging errors drawn from the complete sonar error simulator), and the nominal Gaussian model p_s (red), as well as the histogram density estimate. (c) The cost functions $r_i^2(x)$ obtained by applying Theorem 1 to the sonar ranging error density estimates shown in (b).

	Mean (m)	Median (m)	Max (m)	Standard deviation (m)
Mixture model p_r	.179	.126	1.34	.191
Optimal Gaussian model p_g	.532	.453	1.32	.281
Nominal Gaussian model p_s	3.91	3.29	17.6	2.89

TABLE I
FINAL LANDMARK POSITIONING ERRORS

	Mean (m)	Median (m)	Max (m)	Standard deviation (m)
Mixture model p_r	.237	.168	1.57	.240
Optimal Gaussian model p_g	.521	.419	2.21	.394
Nominal Gaussian model p_s	4.30	3.23	20.9	3.72

TABLE II
FINAL AUV POSITIONING ERRORS

	Mean (sec)	Median (sec)	Max (sec)	Standard deviation (sec)
Mixture model p_r	1.97 E-2	7.93 E-3	.998	9.69 E-2
Optimal Gaussian model p_g	1.24 E-2	6.70 E-3	.997	7.23 E-2
Nominal Gaussian model p_s	1.38 E-2	6.98 E-3	.998	7.79 E-2

TABLE III
ELAPSED COMPUTATION TIMES (PER ITERATION)

by virtue of (15) and (17). 16000 samples were also drawn from the complete (internal receiver and multipath) sonar error simulator, and their sample mean $\mu_g = .999$ and variance $\sigma_g^2 = 3.85$ used to construct the optimal Gaussian approximation p_g to the complete sonar ranging error density via moment-matching. Figure 4 shows the density estimates obtained for the multipath error (Fig. 4(a)) and the complete sonar ranging error (Fig. 4(b)) according to this procedure, together with their histogram density estimates.

The experiment itself consisted of simulating the sensor observations acquired by the AUV as it made two passes over the search pattern shown in Fig. 3. Altogether there were 3629 simulated sensor measurements, of which 2153 were sonar returns from the landmarks and the remaining 1476 were odometry constraints coming from the AUV's IMU and DVL ego-motion estimate. This data set was processed three times, once using the mixture model p_r in (14) in combination with the method of Section II-C to estimate the sonar ranging error, again using the optimal Gaussian model p_g, and once more using the nominal Gaussian model p_s. In all cases, maximum likelihood estimation was performed online using RISE2 with the default settings. Results from this experiment are shown in Fig. 3 and summarized in Tables I – III.

As can be seen from Tables I and II, the error in the AUV and landmark positioning estimates when using the mixture model p_r is approximately half of what it is when using the best possible Gaussian model p_g. By explicitly representing the two cases of line-of-sight and multipath reflection returns, the mixture model p_r is able to correctly exploit the (relatively) high accuracy of the sonar ranging measurements in the former case while allowing for high uncertainty in the ranging measurement in the latter case. This can be seen directly in the cost functions shown in Fig. 4(c): the cost function obtained from p_r closely tracks that of the nominal model p_s near 0 (thus providing strong localization in the case of line-of-sight returns), but attenuates the cost associated with large ranging errors to correctly model the high uncertainty associated with multipath reflection returns. In contrast, the optimal Gaussian

model p_g obtained by moment-matching attempts to "split the difference" between the two sonar ranging modes, and consequently actually assigns much of its probability mass to a region of the state space that is in reality highly unlikely (observe that the mode of p_g, as shown in Fig. 4(b), lies in a region of the state space that the histogram density approximation assigns very little probability mass).

Furthermore, we observe that while the results obtained using the maximum entropy model $\hat{p}_m$ in (16) are already sufficient to show significant improvements over the optimal Gaussian model p_g in this toy example, in reality it should be possible to obtain still better results by better characterizing the multipath error density p_m. As can be seen in Fig. 4(a), the maximum entropy model $\hat{p}_m$ turns out not to be a particularly good representation of p_m. This is actually not terribly surprising: we selected the maximum entropy density not because we expected it to be a particularly good match, but because it is the epistemologically correct default model to use given limited information about the true underlying distribution [8], [22], a fact which is closely related to its being the maximally uncertain distribution subject to the given constraints (in this case, an enforced mean and variance). The adoption of this model is a very conservative approach (similar in spirit to the use of minimax estimators), but may fail to exploit the available data as effectively as would a model which better characterizes the true underlying multipath distribution. Thus, we expect that it should be possible to improve the results given for p_r in Tables I and II by better characterizing p_m. In contrast, the model p_g already provides the *best possible* performance among all Gaussian models. Thus, the experimental results contained herein are themselves conservative estimates of how well the method of Section II-C can perform versus the purely Gaussian assumption.

Finally, we observe that although the mixture model p_r given in (14) is more complicated, and hence more computationally expensive, than the Gaussian models p_g and p_s, this additional computational burden is small compared to the cost of running the entire RISE2 algorithm (cf. Table III). The improvement in performance is well worth the additional expense, as all three of these models are fast enough to run easily in real-time.

V. Related Work

While the method of Section II-C applies quite generally, our research is motivated primarily by the SLAM and bundle adjustment problems; consequently, our discussion will focus on those fields.

Historically, graph-based approaches to solving the full SLAM [14], [23], [24] and bundle-adjustment [6], [25] problems have been formulated under the assumption of mean-zero additive Gaussian noise. This is computationally expedient, as it implies that maximum likelihood estimation is equivalent to minimizing a weighted sum of squared measurement errors; consequently, essentially all publicly available state-of-the-art solvers for both graphical SLAM [9], [11], [12], [20] and bundle adjustment [13] are implemented as sparse least-squares minimization algorithms.

However, it has long been recognized that estimation via quadratic cost minimization performs very poorly when the true sampling distribution may generate gross errors. Several methods have been proposed for addressing this shortcoming.

A standard approach in bundle adjustment is to replace the quadratic cost function with a robust cost function designed to attenuate the ill effects of gross errors (cf. [6, Sec. A6.8] for a list of commonly-used robust cost functions). While these methods have proven to be effective in practice, they often lack a solid theoretical foundation. Many of them are purely heuristic (i.e. they are engineered by hand to have certain outlier attenuation properties, rather than derived by considering the true underlying noise model). These also generally contain a free parameter b that controls the cutoff for attenuation, which must be hand-tuned in a more or less arbitrary fashion, and often at great effort or expense. Furthermore, even those robust cost functions that *do* have solid theoretical underpinnings (e.g. the Huber robust cost function [7]) still do not make as efficient use of the available data as would the use of the maximum likelihood estimator corresponding to the true underlying distribution; they are an intentionally conservative (hence also wasteful) choice.

More recently, attempts have been made to improve the robustness of estimation in localization and full SLAM through the development of models that better capture the true underlying noise process (for example, the formulation of the simulation in Section IV-B was partially motivated by recent work of Prorok *et al.* [19], who achieved improved performance in ultra-wideband localization by better characterizing multipath reflection errors). One notable recent approach that is spiritually similar to our own is due to Olson and Agarwal [16], who have achieved impressive robustness results in SLAM applications by explicitly modeling sensor failure modes such as false loop closures and wheel slippage. Their work goes beyond the simple Gaussian assumption by admitting probability densities that are *mixtures* of Gaussians, and presents an approximate inference method for performing computationally efficient online inference over such models (briefly, at each timestep, each mixture density is approximated by its constituent component that has the greatest likelihood). This method can be expected to work well when the true underlying distribution to be modeled has widely separated modes, each of which can be well-approximated by a Gaussian (and in those cases, it serves as a computationally expedient and easily implemented approximation to the method of Section II-C).

However, there are many distributions that one expects to arise in practical scenarios that do not satisfy these conditions. For example, consider the exponential distribution (the maximum entropy distribution on $[0, \infty)$ with a specified mean). While it is possible to approximate this distribution using a sum of Gaussians, this approach is computationally expensive (since it requires multiple summands in order to obtain a reasonably good approximation), and (more perniciously)

introduces false local maxima that can entrap gradient-based optimizers. Similarly, the analysis of Section IV-A proves that no summation of Gaussians can accurately model a Cauchy distribution, *even in principle*, because such a sum is incapable of capturing the tail behavior.

In contrast with this prior work, our approach provides a simple and mathematically principled method for performing computationally efficient online inference over sparse non-Gaussian distributions in the general case.

VI. Conclusion

In this paper we developed a method for generalizing iSAM and RISE by removing the assumption of Gaussian factors, thereby significantly expanding the class of distributions to which these algorithms can be applied. The generalization is achieved by means of a simple algebraic reduction that under relatively mild conditions (boundedness of each of the factors in the distribution of interest) enables an instance of maximum likelihood estimation to be reduced to an equivalent instance of least-squares minimization that can be solved efficiently online by application of iSAM or RISE. Through this construction we obtain robust, computationally efficient, and *mathematically correct* incremental online maximum likelihood estimators for non-Gaussian distributions over sparse factor graphs.

It should come as no suprise that the use of the mathematically correct maximum likelihood estimator can dramatically improve the quality of the resulting estimate versus what is achievable under the naïve Gaussian assumption. Indeed, for most engineering applications, simple direct maximum likelihood estimation provides a mathematically principled and provably robust solution to the inference problem, *provided that the correct estimator is used*. Our goal in pursuing this research has been to provide a simple, computationally efficient framework that enables the more widespread application of these estimators in practice.

Of course, the use of the mathematically correct maximum likelihood estimator does require some knowledge of what the true underlying distribution actually *is*. In some cases this is known *a priori* through theoretical means (e.g. models that come directly from physics, etc.); otherwise we must attempt to estimate this from data by applying model learning techniques. Conversely, given a model learned from data, one cannot assume *a priori* that this model will be Gaussian; therefore, the effective practical application of model learning requires the availability of inference algorithms that are capable of operating on non-Gaussian distributions. Thus, we see that model learning techniques and non-Gaussian inference methods are naturally complementary technologies. Consequently, we intend to explore approaches to model learning complementary to the generalized iSAM/RISE framework in subsequent research.

Acknowledgments

This work was partially supported by the Office of Naval Research under grants N00014-10-1-0936, N00014-11-1-0688, N00014-12-1-0020 and N00014-12-1-0093, which we gratefully acknowledge.

References

[1] C.R. Blyth. Convolutions of Cauchy distributions. *The American Mathematical Monthly*, 93(8):645–647, 1986.

[2] X. Cheng, H. Shu, Q. Liang, and D. Hung-Chang Du. Silent positioning in underwater acoustic sensor networks. *IEEE Transactions on Vehicular Technology*, 57(3):1756–1766, May 2008.

[3] W.H. DuMouchel. On the asymptotic normality of the maximum-likelihood estimate when sampling from a stable distribution. *The Annals of Statistics*, 1(5):948–957, 1973.

[4] T.S. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall/CRC, 1996.

[5] B.V. Gnedenko and A.N. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, 1968.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[7] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[8] E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, September 1968.

[9] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert. iSAM2: Incremental smoothing and mapping using the Bayes tree. *The International Journal of Robotics Research*, 31:217–236, February 2012.

[10] M. Kaess, A. Ranganathan, and F. Dellaert. iSAM: Incremental smoothing and mapping. *IEEE Trans. Robotics*, 24(6):1365–1378, Dec 2008.

[11] K. Konolige, J. Bowman, J.D. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua. View-based maps. *Intl. J. of Robotics Research*, 29(8):941–957, 2010.

[12] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A general framework for graph optimization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, Shanghai, China, May 2011.

[13] M.I.A. Lourakis and A.A. Argyros. SBA: A software package for generic sparse bundle adjustment. *ACM Trans. Math. Softw.*, 36(1):1–30, March 2009.

[14] F. Lu and E. Milios. Globally consistent range scan alignment for environmental mapping. *Autonomous Robots*, 4:333–349, April 1997.

[15] P. Newman and J. Leonard. Pure range-only sub-sea SLAM. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1921–1926, September 2003.

[16] E. Olson and P. Agarwal. Inference on networks of mixtures for robust robot mapping. In *Robotics: Science and Systems (RSS)*, Sydney, Australia, July 2012.

[17] J. Pitman. *Probability*. Springer-Verlag, New York, 1993.

[18] M.J.D. Powell. A new algorithm for unconstrained optimization. In J. Rosen, O. Mangasarian, and K. Ritter, editors, *Nonlinear Programming*, pages 31–65. Academic Press, 1970.

[19] A. Prorok, L. Gonon, and A. Martinoli. Online model estimation of ultra-wideband TDOA measurements for mobile robot localization. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 807–814, St. Paul, MN, May 2012.

[20] D.M. Rosen, M. Kaess, and J.J. Leonard. An incremental trust-region method for robust online sparse least-squares estimation. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1262–1269, St. Paul, MN, May 2012.

[21] A.A.M. Saleh and R. Valenzuela. A statistical model for indoor multipath propagation. *IEEE Journal on Selected Areas in Communications*, 5(2):128 –137, February 1987.

[22] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, Jan 1980.

[23] S. Thrun. Robotic mapping: A survey. In G. Lakemeyer and B. Nebel, editors, *Exploring Artificial Intelligence in the New Millenium*. Morgan Kaufmann, 2002.

[24] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT Press, Cambridge, MA, 2008.

[25] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 153–177. Springer Berlin / Heidelberg, 2000.