# PARTICLE FILTERING UNDER COMMUNICATIONS CONSTRAINTS

*Alexander T. Ihler*

Donald Bren School of Info. & Comp. Sci.
U.C. Irvine, Irvine, CA 92697-3425

*John W. Fisher III, Alan S. Willsky*

Massachusetts Institute of Technology
Cambridge MA 02139

## ABSTRACT

Particle filtering is often applied to the problem of object tracking under non-Gaussian uncertainty; however, sensor networks frequently require that the implementation be local to the region of interest, eventually forcing the large, sample-based representation to be moved among power-constrained sensors. We consider the problem of successive approximation (i.e., lossy compression) of each sample-based density estimate, in particular exploring the consequences (both theoretical and empirical) of several possible choices of loss function and their interpretation in terms of future errors in inference, justifying their use for measuring approximations in distributed particle filtering.

## 1. INTRODUCTION

The problem of tracking moving objects in intelligent distributed sensor networks has been widely studied. Typically, such networks are comprised of nodes which have some local computational power, the ability to sense their local surroundings, and the ability to communicate with nearby sensors. These sensors operate using a limited, on-board power supply, which must be conserved to prolong sensor lifetime. Usually, the energy costs related to computation are significantly lower than energy costs associated with communication. The disparity in energy costs has led to several approaches which utilize local computational power in order to save communications costs. In these a approaches, a single node is responsible for receiving measurements and combining them to update the state distribution. This is coupled with a strategy of dynamically changing which node is responsible for the state update, usually a sensor which is "close" to the target [1, 2].

Another aspect of tracking problems in these networks is that the measurement models are often nonlinear functions of the state, resulting in non-Gaussian posterior distributions. This has made particle filters and other nonparametric methods attractive alternatives to traditional Kalman filtering approaches for state estimation. In these methods, the state distribution is characterized by a large collection of samples rather than a parametric family.

In combination, these factors necessitate transmission of large, sample-based representations of the state distribution between sensors over time, which can be quite costly. Quantization and compression of the model can mitigate this cost and prolong the operational lifetime of the sensor network. However, for lossy compression, the savings comes at the expense of inference quality due to the transmission of an approximate model.

This research was supported in part by MIT Lincoln Laboratory under Lincoln Program 2209-3023 and by ODDR&E MURI through ARO grant DAAD19-00-0466.

As noted in [1], parametric (e.g., Gaussian) approximations can be communicated efficiently but have limited representational power, while non-parametric (particle-based) approximations can capture arbitrarily complex uncertainty, but at high cost in communications. Ideally, one would like to be able to smoothly trade off the representational power of the model with the cost of transmission, in a way that minimizes the degradation of inference.

One such flexible solution can be found by interpreting the problem as one of lossy encoding of a non-parametric density estimate. In this case, an appropriate measure of "loss" must capture not only the immediate error between the true and approximated distributions, but also the *future* effects of this error. Specifically, the approximation of any particular density has the potential to adversely affect each subsequent estimate of target state. This aspect makes the notion of measuring "loss" for densities a key, nontrivial question.

## 2. PARTICLE FILTERING

Let us begin by considering a nearly canonical problem in sensor networks: that of tracking a moving object as it passes through a field of myopic sensors. We assume a simple Markov process for the underlying dynamics of the object of interest, specified by the transition probability $p(x_t|x_{t-1})$, and at each time step $t \geq 1$ obtain an observation $y_t$ about the state $x_t$, conditionally independent of the other variables given $x_t$. Defining the observation history at time $t$ by the vector $Y_t = [y_1, \dots, y_t]$, the (centralized) filtering problem is to estimate the posterior distribution $p(x_t|Y_t)$ for each $t$; this can be performed via the recursive relationships

$$p(x_t|Y_{t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|Y_{t-1})\, dx_{t-1} \qquad (1)$$

$$p(x_t|Y_t) \propto p(y_t|x_t)p(x_t|Y_{t-1}) \qquad (2)$$

For arbitrary forward dynamics $p(x_t|x_{t-1})$ and likelihood functions $p(y_t|x_t)$, the operations (1)-(2) do not possess closed-form solutions. However, particle filtering [3] can be applied as a Monte Carlo approximation to each step. In particle filtering, each distribution is represented by a set of $N$ weighted samples $\{w_i, \mu_i\}$,

$$\hat{p}(x_t|Y_t) = \sum_i w_i\, \delta(x_t - \mu_i) \qquad (3)$$

and the two operations of Eq. (1)-(2) are performed by sequential resampling and weighting; in the simplest version of particle filtering, $\hat{p}(x_t|Y_{t-1})$ is constructed by drawing samples from $\hat{p}(x_{t-1}|Y_{t-1})$ and then (stochastically) propagating each sample through the forward dynamics $p(x_t|x_{t-1})$; $\hat{p}(x_t|Y_t)$ is then constructed by weighting each sample by $p(y_t|x_t)$ and normalizing $\sum w_i = 1$.

In order to make the notion of approximating the posterior distribution $\hat{p}(x_t|Y_t)$ well-defined, we will additionally assume that it is smooth; in particular, we use a "regularized" particle filter, in which each distribution is represented by a kernel density estimate [4]:

$$\hat{p}(x_t|Y_t) = \sum_i w_i \; K_{\Sigma_i}(x_t - \mu_i) \tag{4}$$

where the kernel function $K$ is a Gaussian distribution with covariance matrix specified by $\Sigma_i$, making $\hat{p}(x_t|Y_t)$ a Gaussian sum.

## 3. TRACKING UNDER COMMUNICATIONS CONSTRAINTS

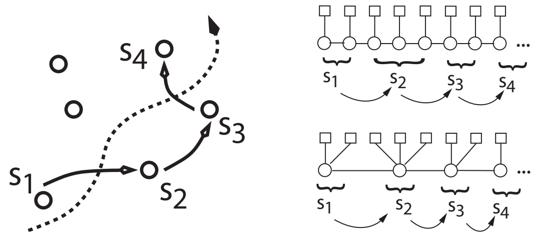
When sensors are myopic (i.e., only observe objects which are nearby to their own location) and constrained by a limited power budget, it is typical to perform the operations of particle filtering at some sensor which is nearby to the object itself, reducing the distance over which sensor observations must be communicated. The sensor in charge of data fusion has been called the "leader" node [5]. At each time $t$, the leader node, along with a few other sensors (perhaps zero), each collect observations (for example, range measurements) and transmit them to the leader, who uses them to update the posterior distribution according to (2).

Because the object is moving, however, the most appropriate leader node is also a function of time. Therefore, at each time $t$, the leader also uses its estimate of the posterior distribution to select a new leader node at time $t+1$. The old leader may, of course, select itself as the new leader; but if not, it must also communicate the current model of the posterior distribution to the new leader [as depicted in Fig. 1(a)]. If the distribution is estimated using a nonparametric representation, the naive cost of this communication can be hundreds or even thousands of times larger than any single measurement communication.

There are any number of possible protocols for selecting when, and to which sensor, one should transfer control; for one example, see [1]. Another, related issue is the decision of which sensors should collect measurements at each time step. However, we shall treat both of these questions as fixed aspects of the leadership protocol, and concentrate on minimizing the communications cost inherent in any given strategy. The fact that the leadership sequence is fixed also means we may ignore the distance-dependent aspect of communications cost, and focus simply on the representation size, for example measured in bits.

## 4. SEQUENTIAL DENSITY APPROXIMATIONS

The problem of sequentially approximating and updating density estimates is in several ways fundamentally different from typical lossy data compression tasks. In essence, these differences are due to the fact that the approximated data is being *used* at the next time step to construct a new density estimate, and that we are interested in minimizing not only the error in the *transmitted* density, but also its effect on subsequent estimates. In other words, we are interested not only in the error introduced in the distribution to reduce communications cost (which is directly calculable by the sender), but *also* (and perhaps more importantly) the error that this difference induces in the updated posterior estimates at each subsequent time step. In many ways, this task is similar to the objective–optimized quantization approach of [6]. To control the resulting inference error, we must understand a particular error's effect in the future.




(a) (b)

**Fig. 1**. (a) Repeated transfer of "leadership", along with the model of state uncertainty, while tracking in an ad-hoc sensor network. (b) Markov chain representation of the sequential state estimation problem; without loss of generality we may assume that the model is transmitted at each time step.

$$\begin{array}{ccccccc}
p^0(x_0) & \rightarrow & p^0(x_1|Y_1) & \rightarrow & p^0(x_2|Y_2) & \rightarrow & \ldots \\
 & & \wr & & & & \\
 & & p^1(x_1|Y_1) & \rightarrow & p^1(x_2|Y_2) & \rightarrow & \ldots \\
 & & & & \wr & & \\
 & & & & p^2(x_2|Y_2) & \rightarrow & \ldots
\end{array}$$

**Fig. 2**. Repeated, sequential approximation in a Markov chain. The true distribution $p^0$ is approximated at time 1; this approximation is then updated and re-approximated at time 2, and so forth.

Of course, the quality of our ideal, particle-based representation is a function of $N$, the number of samples used in (3); however, as we are further approximating the estimated distribution for each transmission, the communications cost is decoupled from $N$, and $N$ is constrained only by computational resources. The (finite) value of $N$ is another source of error at each resampling step of particle filtering. However, for the purposes of this paper, we will assume that $N$ is sufficiently large that the particle-based (internal) representations accurately represent each of the required distributions. In practice, one may wish to estimate the level of error in each resampling step (by, for example, performing the operation twice and comparing the results) to establish a baseline below which the error caused by approximations for communication purposes will go unnoticed.

Let $p^i(x_t|Y_t)$ denote the density of the state $x_t$ at time $t$, given observations $Y_t = [y_1, \ldots, y_t]$ up to time $t$, and subject to approximations at times $1 \ldots i$. So, for example, $p^0(x_0)$ is the original (exact) prior distribution, $p^0(x_1|Y_1)$ is the true posterior at time $t = 1$, $p^0(x_1|Y_1) \propto p(y_1|x) \int p(x_1|x_0)p^0(x_0) \; dx_0$, and $p^1(x_1|Y_1) \approx p_0(x_1|Y_1)$ is the density transmitted at time $t = 1$ to the leader at time $t = 2$. These relationships are also displayed in Fig. 2. As a convenient shorthand, let us define $p_t^i \equiv p^i(x_t|Y_t)$.

Our goal is to show that for certain error measures, we may use the quality of the approximations between $p_t^{t-1}$ and $p_t^t$ to obtain some information about the quality between the true distribution $p_t^0$ and $p_t^t$. The next few sections are concerned with analyzing the (cumulative) effects of a sequence of approximations, as measured by three prospective methods.

This problem is difficult simply because each approximation affects the *true* error (distance from $p^0$) inherent in a later approximation at time $t$, which of course has access only to $p^{t-1}$. Consider, for example, a simple example of a one-dimensional Gaussian distribution with mean zero and standard deviation $\sigma$. Now suppose that one approximates the variance at each time $i$ by $\sigma^i$, such that $\sigma^i/\sigma^{i-1} \approx .96$. If we measure error in terms of the Kullback-Leibler (KL) divergence, we find that the error between $p^{i-1}$ and $p^i$ is always less than $10^{-3}$. However, the error at time $t = 10$ is *not* $10^{-2}$, but more than $10^{-1}$, and this gap increases with $t$.

### 4.1. Error measures

In this work, we explore the relative merits of three possible measures of approximation error; in particular, the maximum log-error (originally applied to this problem in [7]),

$$\mathrm{ML}[p,q] = \max_x |\log p(x) - \log q(x)|\,, \tag{5}$$

the Kullback-Leibler (KL) divergence,

$$\mathrm{KL}[p,q] = \int p(x) \log \frac{p(x)}{q(x)}\, dx, \tag{6}$$

and the integrated absolute, or $\mathrm{L}_1$, error

$$\mathrm{L}_1[p,q] = \int |p(x) - q(x)|\ dx. \tag{7}$$

These three measures are listed in order of decreasing strictness, in the sense that each bounds the next—for any distributions $p, q$, we have

$$(\mathrm{L}_1[p,q])^2 \le 2\,\mathrm{KL}[p,q] \le 2\,\mathrm{ML}[p,q], \tag{8}$$

but no converse inequalities hold for arbitrary $p, q$. As one consequence, we shall see that ML has the best theoretical properties, but may be an overly strict assessment of error; using a more relaxed criteria such as KL or $\mathrm{L}_1$ can improve performance, but has weaker theoretical guarantees.

### 4.2. Maximum Log-Error

It turns out to be straightforward to analyze the resulting error of a sequence of approximations under the maximum log-error (ML) measure applied in [7]. In particular, it is closely related to a logarithmic "dynamic range"

$$\mathrm{DR}[p,q] = \max_x \min_\alpha |\log p(x) - \log q(x) + \alpha|$$

in that for any distributions $p, q$

$$\mathrm{DR}[p,q] \le \mathrm{ML}[p,q] \le 2\,\mathrm{DR}[p,q]. \tag{9}$$

The measure DR, in turn, has two properties which make analysis straightforward—it satisfies the triangle inequality

$$\mathrm{DR}[p_t^0, p_t^2] \le \mathrm{DR}[p_t^0, p_t^1] + \mathrm{DR}[p_t^1, p_t^2] \tag{10}$$

and also bounds the error after incorporating (arbitrary) new information between time $i+1$ and some later time $t$:

$$\mathrm{DR}[p_t^i, p_t^{i+1}] \le \mathrm{DR}[p_{i+1}^i, p_{i+1}^{i+1}] \tag{11}$$

for any $t > i$. Applying these three identities recursively to the total error, we see that

$$\mathrm{ML}[p_t^0, p_t^t] \le 2 \sum_{i=0}^{t-1} \mathrm{ML}[p_{i+1}^i, p_{i+1}^{i+1}] \tag{12}$$

each term of which is computable at time $i+1$ when the approximation is made.

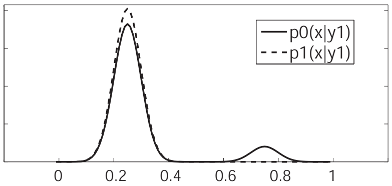



**Fig. 3**. Given two distributions $p^0(x_1|y_1)$ and $p^1(x_1|y_1)$ with relatively small $\mathrm{L}_1$ error, a new observation $y_2$ can either increase (if $y_2 = 1$) or decrease (if $y_2 = 0$) the $\mathrm{L}_1$ error between $p^0(x_1|y_1, y_2)$ and $p^1(x_1|y_1, y_2)$, though one can bound the error in expectation over $y_2$; see text for details.

The bound (12) is a strong theoretical statement about the total error which could arise from any given approximation. In this sense, the maximum log-error is a reasonable quantity to trade off with communications costs. However, it also has a number of drawbacks. First, it is an extremely strict criterion—for example, it heavily penalizes errors in regions with low probability. Since, by definition, the true state is unlikely to lie in such a region, intuitively one would prefer an approximation which better in high-probability regions, at some cost to low-probability ones.

This non-intuitive emphasis is, in some sense, a result of the "max" operation applied to a choice of point-wise error ($\log p/q$); as we shall see, this is necessary to bound the effect of errors given an arbitrary set of measurements $Y_t$. However, more common measures of difference between two distributions measure the total ($\mathrm{L}_1$) or weighted average (KL); we examine these measures next.

### 4.3. $\mathrm{L}_1$ Error

One common measure of the difference between two densities is the integrated absolute (or $\mathrm{L}_1$) error [8], given by (7). It is easy to show that the triangle inequality still holds:

$$\mathrm{L}_1[p_t^0, p_t^2] \le \mathrm{L}_1[p_t^0, p_t^1] + \mathrm{L}_1[p_t^1, p_t^2].$$

The incorporation of new information, however, has a more complex effect. It turns out that it is impossible to say anything in general about the $\mathrm{L}_1$ error with respect to a *particular* sequence of measurements $[y_1, \ldots, y_t]$. To see this, consider the distributions $p^0(x_1|Y_1)$ and $p^1(x_1|Y_1)$ depicted in Fig. 3 (which have an $\mathrm{L}_1$ difference of .2) and suppose

$$x_2 = x_1 \qquad \text{and} \qquad y_2 = \begin{cases} 1 & x_2 > .5 \\ 0 & x_2 \le .5 \end{cases}$$

Now, if we observe $y_2 = 0$ the resulting posterior distributions $p^0(x_2|Y_2)$ and $p^1(x_2|Y_2)$ have essentially zero $\mathrm{L}_1$ error; but, if we observe $y_2 = 1$ the two distributions are radically different and have an $\mathrm{L}_1$ difference of nearly 2 (the maximum possible). Thus the initial $\mathrm{L}_1$ error tells us little about the resulting error for an arbitrary observation $y_2$. However, given that $p^0$ is the true posterior, our probability of observing the catastrophic $y_2 = 1$ is quite low. We make this notion precise by considering the $\mathrm{L}_1$ error *in expectation* over the observations $Y_t$. This allows the possibility of low-probability, catastrophic errors (such as described) and concentrates on estimating the "average" resulting level of error.

One may relate the expected error which the approximation made at time $i+1$ incurs at a later time $t$ to the error at time $i+1$

by the bound

$$E[\mathrm{L}_1[p_t^i, p_t^{i+1}]] \leq 2E[K(i,t)\,\mathrm{L}_1[p_{i+1}^i, p_{i+1}^{i+1}]]$$

where the expectation $E[\cdot]$ is over the observations $Y_t$, and

$$K(i,t) = \max_{Y_t|Y_{i+1}} p^0(Y_t|Y_{i+1})/p^i(Y_t|Y_{i+1}).$$

Although the quantity $K(i,t)$ is difficult to calculate, it can be bounded using the ML error measure:

$$E[\mathrm{L}_1[p_t^i, p_t^{i+1}]] \leq 2E[\exp(\mathrm{ML}[p_i^0, p_i^i])\ \mathrm{L}_1[p_{i+1}^i, p_{i+1}^{i+1}]]$$

This bound has some interesting interpretations. First, it suggests that there is in some sense a "dual" cost for each approximation—the current effect of an error, as measured by $\mathrm{L}_1$, and its possible future effects, measured by ML. We know that the ML-based error bound is conservative; thus we might consider optimizing for some finite window of time, after which we hope that dynamics and redundant information have negated any further influence of errors. The limits of such a method are intuitive—for a window of length 1, $\mathrm{ML} = 0$ and the error is at most twice the $L_1$ cost between $p^0$ and $p^1$; for large window lengths, the exponential dependence implies that we should be primarily concerned with minimizing the ML error.

In practice, we find that often the simple assumption of taking $K(i,t) = 1$, equivalent to optimizing over a length-one window, works quite well. As an approximation, it is perhaps justified given the derived bounds and the relatively complex ways the various errors must compound in order to become super-additive.

### 4.4. Kullback-Leibler Divergence

Kullback-Leibler (KL) divergence is another common measure of the difference between two distributions. However, it lacks a version of the triangle inequality similar to (10). One may, however, claim the following:

$$\mathrm{KL}[p_t^0, p_t^2] \leq \mathrm{KL}[p_t^0, p_t^1] + C(1,t)\,\mathrm{KL}[p_t^1, p_t^2]$$

where the amplification constant

$$C(i,t) = \max_{x_t} p^0(x_t|Y_t)/p^i(x_t|Y_t)$$

captures the fact that subsequent errors may compound upon each other in a super-additive way, as was the case for the Gaussian example given in Sec. 4.

As with the $\mathrm{L}_1$ measure, one cannot bound the error at a future time $t$ for arbitrary observation sequences $Y_t$. However, one can again bound the future KL-divergence *in expectation* over the observations. Specifically, it is well known [9] that

$$E[\mathrm{KL}[p_t^0, p_t^1]] \leq E[\mathrm{KL}[p_1^0, p_1^1]]$$

where the expectation $E[\cdot]$ is again taken over the $Y_t$.

By using these two properties, one may again demonstrate a bound similar to that shown for $\mathrm{L}_1$:

$$E[\mathrm{KL}[p_t^i, p_t^{i+1}]] \leq E[\exp(\mathrm{ML}[p_i^0, p_i^i])\ \mathrm{KL}[p_{i+1}^i, p_{i+1}^{i+1}]]$$

which enables one to apply the same kind of "dual-cost" interpretations, as well as simple approximations such as that given by taking $C(i,t) = 1$.

## 5. ESTIMATING THE ERROR MEASURES

We next consider how one may apply these bounds to control a system's inference error. The first step is to be able to either evaluate or estimate the error measures themselves. It will turn out (see Sec. 6) that we will be primarily concerned with estimating the error between a kernel density estimate $p$ as in (4) and a *single* Gaussian component $q$; to do so we use simple plug-in estimates:

$$\mathrm{ML}[p,q] \approx \max_{\mu_i} |\log p(\mu_i) - \log q(\mu_i)|$$
$$\mathrm{KL}[p,q] \approx \sum w_i \log p(\mu_i) - \log q(\mu_i)$$
$$\mathrm{L}_1[p,q] \approx \sum w_i\,|1 - q(\mu_i)/p(\mu_i)|$$

In controlling average-case behavior, one subtlety which arises is that the bounds are in terms of the *expected* error at each time step, while we have access to only one realization (the actual distribution to be transmitted). If we are only interested in minimizing error subject to some per-message communications constraint, this makes little difference—we simply minimize each distribution's error, and attain some (unknown) level of expected error. However, if we instead have some total energy budget, so that it is useful to send less information for "simpler" distributions, this opens the door to more complicated strategies in which each sensor attempts only to make the *average* error introduced meet some threshold. For simplicity, we will concentrate on minimizing error for some fixed, per-transmission cost; extending to more holistic strategies comprises an interesting open area of research.

## 6. OPTIMIZATION

Given a measure of error (or loss) for an approximate representation, we now consider how to minimize this loss subject to a given communications budget. It is convenient to select a method which can be adapted to any of the three loss measures, such as the KD-tree based approximation described in [7].

A "k-dimensional" tree, or KD-tree, is a data structure for rapidly performing locality-based computations on large sets of continuous-valued points. Specifically, a KD-tree is a binary tree structure whose leaf elements each store one point $\mu_i$, or more generally (in our application) the point $\mu_i$, weight $w_i$, and associated covariance $\Sigma_i$. Internal (non-leaf) nodes store sufficient statistics for the data represented by their children, which enable fast exact or approximate computations.

We apply KD-trees to define a hierarchy of Gaussian mixture approximations to a sample-based density estimate $p(x)$. The sufficient statistics stored at each node are thus the mean and covariance of the Gaussian sum defined by the node's children, along with a weight $w_i$ representing the sum of the weights of its associated leaf nodes; all three quantities can be easily computed recursively for each node. Any "cut-set" $S$ (set $S$ which contains one and only one ancestor of each leaf node) can then be used to define a Gaussian sum approximation $q_S$ to the original kernel density estimate $p$.

To trade off communications and error, we first define a communications cost $C(\cdot)$ for the mixture component of each node, such that the cost of any node is less than the sum of the costs associated with its two children. One example cost structure is given in [7], corresponding to the total cost in bits required to describe the mixture component using an encoder defined on the same KD-tree. A simpler example is given by assigning unit cost to each mixture component, which corresponds to the size of a naive representation of each component.

We also create an upper bound $B(S)$ on the error associated with any cut-set $S$ which decomposes into a simple operation (sum or max) on the error $B(s)$ estimated for each node $s \in S$ individually, between the Gaussian approximation stored at that node (denoted $q_s$) and the kernel density estimate defined by its descendant leaf nodes (denoted $p_s$). Such a bound is easy to obtain for all three measures considered here:

$$\begin{aligned} \mathrm{ML}[p, q_S] &\le B_{\mathrm{ML}}(S) = \max_{s \in S} \mathrm{ML}[p_s, q_s] \\ \mathrm{KL}[p, q_S] &\le B_{\mathrm{KL}}(S) = \sum_{s \in S} w_s \ \mathrm{KL}[p_s, q_s] \\ \mathrm{L}_1[p, q_S] &\le B_{\mathrm{L}_1}(S) = \sum_{s \in S} w_s \ \mathrm{L}_1[p_s, q_s] \end{aligned} \tag{13}$$

We then greedily optimize over the KD-tree structure to minimize the bound $B(S)$ subject to a threshold on $C(S)$; see Alg. 1.

Initialize $S^* = S = \{1\}$.
Compute node 1's cost $C(1)$, and error $B(1)$ via (13).
While $\sum_{s \in S} C(s) \le C_{max}$,
- Find $\bar{s} = \arg\max_s B(s)$.
- Exclude $\bar{s}$ and include its children: $S = S \setminus \bar{s} \cup \bar{s}_L \cup \bar{s}_R$
- For left & right child nodes $\bar{s}_L$, $\bar{s}_R$, compute errors $B(\bar{s}_L)$, $B(\bar{s}_R)$ and costs $C(\bar{s}_L)$, $C(\bar{s}_R)$.
- If $B(S) < B(S^*)$, set $S^* = S$.

Return $q_{S^*}$, the density associated with the set $S^*$.

Algorithm 1: Greedy algorithm for approximating a kernel density estimate subject to maximum cost $C_{max}$ by optimizing over Gaussian mixtures defined by a KD-tree.

## 7. SIMULATIONS

We next examine the implications of our analysis using simulations of target tracking in a distributed sensor network. We assume a simple two-dimensional state for the target, with dynamics

$$x_{t+1} = x_t + r_0 \left( [\cos\theta_t; \sin\theta_t] + u_t \right)$$

where $\theta_t \sim U[-\frac{\pi}{4}, \frac{\pi}{4}]$ and $u_t \sim U[0, .1]$. An example target trajectory is shown in Fig. 4. Also shown are a number of sensors scattered in the same region; we indicate the leader node at each time step by drawing a dashed line from the true target position at each time $t$ (circles) to the associated leader node at that time (diamonds). At each time $t$, only the leader node obtains an observation of the target position, given by

$$y_t = 1/(\|s_t - x_t\| + .05) + v_t$$

where $s_t$ is the position of the leader node, $x_t$ that of the target, and $v_t \sim \mathcal{N}(0, .05)$, giving a (saturating) measurement of the inverse range corrupted by Gaussian noise. We use $N = 500$ samples for the particle filter representation at each sensor, and force the leader node to change at each time step and the posterior distribution to be approximated using fewer than $M$ mixture components. Fig. 5 shows an example of the (correct) particle-based posterior estimates at times $t = \{2, 5, 8\}$.

Using this framework, we can assess the performance of our bounds on estimation quality. We consider the average-case behavior of each measure, as assumed by the KL and $\mathrm{L}_1$ bounds;

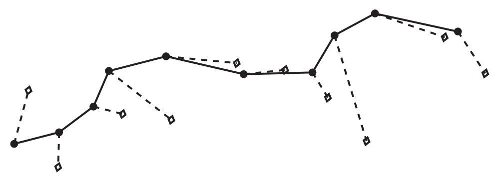


**Fig. 4**. A field of sensors (diamonds), along with the location of the tracked object at each time step (solid, circles), and the leader node measuring range at each time (dashed lines).

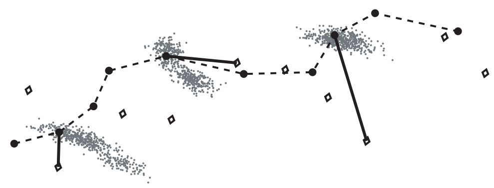


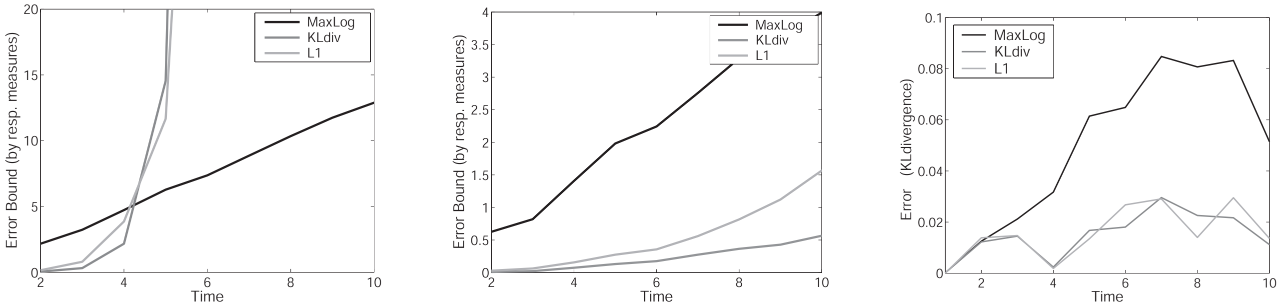
**Fig. 5**. Three example particle-based posterior distributions for the same realization as Fig. 4, at time $t = 2, 5, 8$.

for ML, by linearity of expectation the average ML errors at each time step also give a bound on the average total ML error.

Optimizing over the KD-tree structure to minimize each type of error under the constraint that the total number of components is fewer than $M = 5$, each bound obtained is shown in Fig. 6(a). Note that these bounds are with respect to their *own* error measure (so that the ML curve bounds the average ML error at each time step, while the KL curve bounds the average KL error). To interpret the curves as bounds on some single, common error measure (such as $\mathrm{L}_1$) one should apply the inequalities (8). While the short-term error guarantees for KL and $\mathrm{L}_1$ are better than that for ML, it rapidly becomes difficult to say anything of value, while the ML bound continues to grow linearly.

If we increase the quality of each approximation by increasing $M$ to 10 components, we see that all three bounds improve [Fig. 6(b)]. For the ML error measure, this takes the form of a decreased (linear) rate of growth; however, it markedly improves the utility of the KL and $\mathrm{L}_1$ bounds. While these are still essentially exponential, the improvement in quality significantly slows the rate of growth, making the bounds potentially much more useful than the ML-based bound over the time period under consideration. Interestingly, the KL-based optimization's bound appears to grow more slowly than that of the $\mathrm{L}_1$-based optimization. This is perhaps because it is the (only implicitly controlled) ML error of each approximation which dominates the exponential rate; the KL measure is "closer" to ML, and thus minimizing KL does a better job of implicitly lowering the ML error as well. One may be able to improve the rate still further by explicitly trading off ML versus KL or $\mathrm{L}_1$ error as a dual cost criterion over some fixed window length, or with some discount factor for the future.

These bounds are still typically quite large compared to the actual mean performance. Again approximating with $M = 5$ components, we can look at the increase in *actual* KL-divergence resulting at each time step. To do so, we ran 500 Monte Carlo trials of our system, estimating the posterior at each time step using two high-quality ($N = 2000$) particle filters along with a communications-constrained particle filter for each approximation method. We then compared the average KL-divergence between each posterior and the estimate of the first reference filter, less




(a) (b) (c)

**Fig. 6**. (a) Error bounds at each time step resulting from a KD-tree based optimization for $\mathrm{ML}$, $\mathrm{KL}$, and $\mathrm{L}_1$ error subject to a maximum of 5 final mixture components. (b) Error bounds for $M = 10$ mixture components. Notice the change in scale from (a)—improving the quality of each approximation increases the time scale on which the $KL$ and $L_1$-based bounds are useful. (c) Increase in actual error (measured by the KL-divergence) due to density approximation for $M = 5$.

the KL-divergence between the two references (which can be attributed to the stochastic sampling error for $N = 2000$ samples). The average error increase, as a function of time (and thus number of approximations) is shown in Fig. 6(c). All three error measures perform *much* better than the bounds, while the $\mathrm{ML}$-optimized method seems to perform slightly less well than the KL- or $\mathrm{L}_1$-optimized distributions. This again indicates the fact that, while optimizing for worst-case performance leads to better guarantees in the future, optimizing for shorter term performance in expectation can often do better on average.

## 8. CONCLUSION

Particle filtering is a common and effective tool for tracking problems involving non-Gaussian observation likelihoods and complex uncertainty. However, the number of particles required to be effective is often quite high, while local implementations in sensor networks must operate under communications constraints which limit the number of transmitted particles.

To overcome this, one can perform lossy encoding of the *distribution* represented by the particles, regenerating a nearly equivalent set of particles at the receiver. This leads to two primary questions—how to measure the loss, or error, in the communicated distribution, and how to effectively trade off this loss for compactness in the representation. Focusing on the former question, we argued that a good measure of error controls not only the immediate difference between the distributions, but also (and just as importantly) the errors which *could* arise after both incorporating subsequent information and performing later approximations. Examining three possible measures of error—the maximum log-error ($\mathrm{ML}$), the Kullback-Leibler divergence ($\mathrm{KL}$), and the $\mathrm{L}_1$ error—we showed that each may be interpreted as leading to an upper bound on worst- or average-case error arising in the future. We compared the behavior of each measure in a simulated sensor network under communications constraints.

Empirically, we found that the measure with the best theoretical behavior ($\mathrm{ML}$) is often too strict, overestimating the error likely to be observed. Less strict, average-case measures such as $\mathrm{KL}$ and $\mathrm{L}_1$ are better short-term predictors of performance, but come with only loose guarantees far in the future. The better empirical performance of $\mathrm{KL}$ and $\mathrm{L}_1$ can perhaps be explained in terms of some discount factor; theoretical understanding of this phenomenon is a subject of ongoing research. Although all three bounds appear loose empirically, they provide some justification for sequential density approximation under each metric.

This is only a first step towards understanding the role of approximations in communication–constrained inference, highlighting the importance of considering the appropriate error measure and providing some guidance in the form of upper bounds. Improving these bounds, or replacing them with accurate estimates of later error, is an important direction for future research.

## 9. REFERENCES

[1] F. Zhao, J. Shin, and J. Reich, "Information-driven dynamic sensor collaboration for tracking applications," *IEEE Signal Proc. Mag.*, vol. 19, no. 2, pp. 61–72, Mar. 2002.

[2] J. J. Liu, J. Liu, M. Chu, J. E. Reich, and F. Zhao, "Distributed state representation for tracking problems in sensor networks," in *IPSN*, 2004, pp. 234–242.

[3] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, 2001.

[4] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

[5] F. Zhao, J. Liu, J. Liu, L. Guibas, and J. Reich, "Collaborative signal and information processing: An information-directed approach," *Proc. IEEE*, vol. 91, no. 8, pp. 1199–1209, Aug. 2003.

[6] L. Vasudevan, A. Ortega, and U. Mitra, "Application–optimized quantizers for time delay estimation in sensor networks," in *IEEE Work. Stat. Sig. Proc.*, 2003.

[7] A. T. Ihler, J. W. Fisher III, and A. S. Willsky, "Communication-constrained inference," Tech. Rep. 2601, MIT, Laboratory for Information and Decision Systems, 2004.

[8] Luc Devroye, *A Course in Density Estimation*, vol. 14 of *Progress in Probability and Statistics*, Birkhauser, Boston, 1987.

[9] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.