Chapter 1

An approximate likelihood method for estimating the static parameters in multi-target tracking models

Sumeetpal S. Singh[1] *Nick Whiteley*[2] *and Simon Godsill*[3]

1.1 Introduction

Target–tracking problems involve the on-line estimation of the state vector of an object under surveillance, called a target, that is changing over time. The state of the target at time n, denoted X_n, is a vector in $E_1 \subset \mathbf{R}^{d_1}$ and contains its kinematic characteristics, e.g. the target's position and velocity. Typically only noise corrupted measurements of the state of the object under surveillance are available. Specifically, the observation at time n, denoted Y_n, is a vector in $E_2 \subset \mathbf{R}^{d_2}$ and is a noisy measurement of the target's state as acquired by a sensor, e.g. radar. The statistical model most commonly used for the sequence of random variables $\{(X_n, Y_{n+1})\}_{n\geq 0}$ is the hidden Markov model (HMM):

$$X_0 \sim \mu^\theta(\cdot), \quad X_n|(X_{n-1} = x_{n-1}) \sim f^\theta(\cdot|x_{n-1}), \quad n \geq 1, \tag{1.1}$$

$$Y_n|X_n = x_n \sim g^\theta(\cdot|x_n), \quad n \geq 1. \tag{1.2}$$

The superscript θ on these densities (as well as on all densities introduced subsequently), denotes the dependency of the model on a vector of parameters θ. We will assume a parameterization such that $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$. When the target first appears in the surveillance region, its initial state is distributed according to the probability density μ^θ on E_1. The change in its state vector from time $n-1$ to n is determined by the Markov transition density $f^\theta(\cdot|x_{n-1})$. Furthermore, the observation generated at time n is a function of the target's state at time n and noise, or equivalently generated according to the probability density $g^\theta(\cdot|x_n)$ on E_2, and is conditionally independent of previously generated observations and state values. This model is general enough to describe the evolution of the target and the observations it generates in many applications; see Bar-Shalom and Fortmann (1964), Mahler (2007).

This chapter is concerned with the more complex and practically significant problem of tracking multiple targets simultaneously. In this case the state and observation at each time are *random finite sets* (Mahler (2007)):

$$\mathbf{X}_n = \{X_{n,1}, X_{n,2}, \ldots, X_{n,K_n}\}, \quad \mathbf{Y}_n = \{Y_{n,1}, Y_{n,2}, \ldots, Y_{n,M_n}\}, \quad n \geq 1. \tag{1.3}$$

[1]Signal Processing Laboratory, Department of Engineering, University of Cambridge
[2]Statistics Group, Department of Mathematics, University of Bristol
[3]Signal Processing Laboratory, Department of Engineering, University of Cambridge

Each *element* of $\mathbf{X}_n$ is the state of an individual target. The number of targets K_n under surveillance changes over time due to targets entering and leaving the surveillance region. Some of the existing targets may not be detected by the sensor and a set of false measurements of unknown number are also recorded due to non-target generated measurements. For example, if the sensor is radar, reflections can be generated by fixed features of the landscape. These processes give rise to the measurement set $\mathbf{Y}_n$. (Note its cardinality M_n changes with time.) An added complication usually encountered in applications is that it is not known which observations arise from which targets (if any). The aim in multi-target tracking is to estimate, at each time step, the time-varying state set from the entire history of observation sets received until that time. The task of calibrating the multi-target tracking model is also an important problem faced by the practitioner. In the multi-target model θ includes both the parameters of the individual target model (1.1)-(1.2) and parameters related to the surveillance environment. For example, θ may contain the variance of the noise that corrupts the sensor measurements, the parameter of the distribution of false measurements, etc. In this chapter, in order to estimate the model parameters from the data, an approximate likelihood function is devised and then maximised. Before describing this method, it is necessary to specify the multi-target tracking problem a little more precisely.

The state $\mathbf{X}_n$ evolves to $\mathbf{X}_{n+1}$ in a Markovian fashion by a process of thinning (targets leaving the surveillance region), displacement (Markov motion of remaining individual targets) and augmentation of new points which correspond to new targets entering the surveillance region. The motion of each target that has not left the surveillance region occurs precisely according to (1.1). When a new target is introduced, its initial state is drawn according to the probability density μ^θ in (1.1). If more than one new target is introduced, then the initial states are sampled from μ^θ independently. The observed process is generated from the hidden process through the same mechanisms of thinning, displacement and augmentation with false measurements. (See Section 1.2 for more details.) Displacement here implies that the individual targets, if they generate an observation at time n, do so in accordance with the stated model in (1.2). Mathematically $\mathbf{X}_n$ is a spatial Point Process (PP) on E_1 where E_1 is the state-space of a single target. Likewise, $\mathbf{Y}_n$ is a spatial PP on E_2 where E_2 is the observation space in the single target tracking problem. False measurements and birth of new targets are, for example, assumed to be independent spatial Poisson processes. Let $\mathbf{y}_{1:n} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote the realization of observations received from time 1 to n. (Here $\mathbf{y}_i$ denotes the realization of $\mathbf{Y}_i$.) It is possible to estimate the number of targets and their individual state values from the conditional distribution of $\mathbf{X}_n$ given $\mathbf{Y}_{1:n} = \mathbf{y}_{1:n}$, denoted $p(\mathbf{x}_n|\mathbf{y}_{1:n})$. Due to the need to process the data online, i.e. to update the estimates every time a new observation is received, the sequence of marginal distributions $\{p(\mathbf{x}_n|\mathbf{y}_{1:n})\}_{n\geq 0}$ is often sought after. For each n, $p(\mathbf{x}_n|\mathbf{y}_{1:n})$ has support on the disjoint union $\biguplus_{k\geq 0} E_1^k$ and does not admit an analytic characterisation in general.

Multi-target tracking has long been a focus of research in the Engineering literature, primarily driven by surveillance applications, and now a standard set of tools exist for the analysis of such problems; see for example Blackman and Popoli (1999), Mahler (2007). The most popular algorithm for target tracking is the Multiple Hypothesis Tracking (MHT) algorithm of Reid (1979); see also Blackman and Popoli (1999), Mahler (2007). It is possible to enlarge the model (1.3) to include the (unobserved) associations of the observations to hidden targets. For simple sensor and individual target motion models, the posterior distribution of the unobserved targets and associations admits an analytic characterisation. Furthermore, this pos-

terior distribution can be marginalized to obtain a posterior distribution over the associations only. Even so, the support of this marginal distribution is too large for it to be stored in practice and approximations are made. The most popular approach to date (in the surveillance literature) is to approximate the posterior by retaining only its dominant modes. A popular sub-optimal search algorithm to locate the modes is the MHT algorithm. As more data are gathered over time, which is characteristic of surveillance, the dimension of the support of the posterior distribution of associations increases and searching for the modes exhaustively is not possible. There is a large volume of work dedicated to the computational challenges of this task, i.e., how to implement the search sequentially in time and direct it towards "good" candidates, and how to store the result efficiently; see Blackman and Popoli (1999). It is fair to say that the MHT is complicated to implement, in fact, far more complicated than the algorithms in this work.

Recently, the MHT algorithm has been extended by Storlie et al. (2009) to simultaneously estimate the parameters of the multi–target model. A full Bayesian approach for estimating the model parameters using Markov Chain Monte Carlo was presented in Yoon and Singh (2008) for a simplified model which assumes the individual targets have linear Gaussian dynamics and similar Gaussian assumptions hold for the observations they generate. This Gaussian scenario is highly restrictive and cannot handle non-linear sensors, e.g. bearings measurements.

We now introduce the specific technique we use to construct an approximation of the marginal likelihood of observations. When the unknown number of targets and their states at time 0 is considered to be a realization of a Poisson PP, then it follows that for $n = 1, 2, ...$, the law of $\mathbf{X}_n$ is also Poisson (see section 1.2 for a precise statement of this result including the modeling assumptions). The problem of estimating the number of targets and their individual state values at time n given the observations $Y_{1:n}$ is then greatly simplified if $\mathbf{X}_n$ given $\mathbf{Y}_{1:n}$ can be closely approximated as a Poisson PP. In the tracking literature this problem was studied by Mahler (2003). Mahler derived an expression relating the *intensity* (or the *first moment*) of the conditional distribution of $\mathbf{X}_1$ given $\mathbf{y}_1$ to that of the prior of $\mathbf{X}_1$. The Poisson PP is completely characterised by its first moment and it can be shown that the problem of finding the best Poisson approximation to the conditional distribution of $\mathbf{X}_1$ given $\mathbf{y}_1$ is equivalent to the problem of characterising its intensity; see Mahler (2003). In addition, for the same hidden process dynamic model stated above, Mahler also derived the intensity of the conditional distribution of $\mathbf{X}_2$ given $\mathbf{y}_1$. These results were combined to yield a filter that propagates the intensity of the sequence of conditional densities $\{p(\mathbf{x}_n|\mathbf{y}_{1:n})\}_{n\geq 0}$ and is known in the tracking literature as the Probability Hypothesis Density (PHD) filter. Detailed numerical studies using SMC approximations by Vo et al. (2003, 2005) (and references therein), as well as Gaussian approximations by Vo and Ma (2006) to the PHD filter have since demonstrated its potential as a new approach to multi-target tracking. We are not aware of any study that has specifically characterised the error incurred in approximating $\mathbf{X}_n$ given $\mathbf{y}_{1:n}$ with a Poisson PP. However, in various studies, the merit of the approximation has been confirmed by comparing the estimated intensity with the true number of targets in synthetic numerical examples. In particular, the estimates of the number of targets and their states extracted from the propagated intensity are reasonably accurate even for difficult tracking scenarios. Recent non-target tracking applications of the PHD filter include map building in robotics by Mullane et al. (2008) and tracking sinusoidal components in audio by Clark et al. (2007).

Motivated by this, we explore the use of the same Poisson approximation tech-

nique to derive an approximation of the marginal likelihood of the observed data, i.e. $p(\mathbf{y}_{1:T})$ where T is the length of the data record. The estimate of the model parameters are then taken to be the maximizing argument (with respect to the model parameters) of this approximation of the true marginal likelihood. A gradient ascent algorithm is used to find the model parameters that maximise the likelihood. Although the approximate likelihood function is not computable exactly, it and its gradient may be evaluated using SMC. The approximate likelihood function is "characterised" by sequence of a non-negative functions on E_1, and not the space $\biguplus_{k\geq 0} E_1^k$, and even a simple SMC implementation can be reasonable efficient. (See Section 1.4 for details.) We demonstrate in numerical examples that the approximation to the true likelihood is reasonable as it allows us to learn the static parameters from the data. Even for initialization values very far from the true model parameters, the gradient algorithm is able to converge to a vicinity of the true model parameters.

The remainder of this chapter is structured as follows. The multi-target statistical model is defined in section 1.2. A review of the PHD filter is presented in section 1.3 along with several results from Singh et al. (2009) which are needed to construct the approximation of the likelihood. The approximation of the marginal likelihood of the multi-target tracking model is detailed in section 1.4. Section 1.5 describes a SMC algorithm to evaluate this approximate likelihood and its gradient with respect to the model parameters. A simulation study which empirically assess the performance of the method is presented in section 1.7.3.

1.2 The Multi-target Model

The Poisson PP features prominently in this model and we refer the reader to Kingman (1993) for an introduction to this process. To simplify the notation, only the core ingredients of the multi-target statistical model have their the dependence on θ made explicit. All other derived quantities, while also dependent on θ, have θ omitted from their symbols.

Consider the process of unobserved points $\mathbf{X}_n = X_{n,1:K_n}$, where each element of $\mathbf{X}_n$ corresponds to one target and is a random point in the space E_1. $\mathbf{X}_n$ evolves to $\mathbf{X}_{n+1}$ as follows. With probability $p_S^\theta(x)$, each point of $\mathbf{X}_n$ survives and is displaced according to the Markov transition density on E_1, $f^\theta(x_{n+1}|x_n)$, introduced in (1.1). The random deletion and Markov motion happens independently for each point in $\mathbf{X}_n$. In addition to the surviving points of $\mathbf{X}_n$, new points are "born" from a Poisson process with intensity function $\gamma^\theta(x)$. Denote by $\mathbf{X}_{n+1}$ the PP on E_1 defined by the superposition of the surviving and mutated points of $\mathbf{X}_n$ and the newly born points. At initialisation, $\mathbf{X}_0$ consists only of "birth" points. Simulation from this Poisson model can be achieved by first sampling the cardinality according to the discrete Poisson distribution with parameter value equal to the total mass of the intensity function. The location of the points themselves are then sampled i.i.d. from the normalised intensity function. In the context of the model (1.1), the initial state of each new target will be drawn from the probability density $\gamma^\theta(x)/\int \gamma^\theta(x')dx'$.

The points of $\mathbf{X}_{n+1}$ are observed through the following model. With probability $p_D^\theta(x)$, each point of $\mathbf{X}_{n+1}$, e.g. $x_{n+1,j}$, $j \in \{1, 2, ..., K_{n+1}\}$, generates a noisy observation in the observation space E_2 through the density $g^\theta(y|x_{n+1,j})$. This happens independently for each point of $\mathbf{X}_{n+1}$. Let $\widehat{\mathbf{Y}}_{n+1}$ denote the PP of observations originating from $\mathbf{X}_{n+1}$. In addition to these detected points, false measurements (or clutter points) are generated from an independent Poisson process on E_2 with intensity function $\kappa^\theta(y)$. Denote by $\mathbf{Y}_{n+1}$ the superposition of $\widehat{\mathbf{Y}}_{n+1}$ and these false

measurements, and a realization of $\mathbf{Y}_{n+1} = Y_{n+1,1:M_{n+1}}$ by $\mathbf{y}_{n+1} = y_{n+1,1:m_{n+1}}$.

1.3 A Review of the PHD Filter

This section presents an alternative derivation of the PHD filter which was proposed by Singh et al. (2009). The foundation of the PHD filter is a solution to a simplified inference task, which is to characterise the posterior distribution of a hidden Poisson PP $\mathbf{X}$ given observations $\mathbf{Y}$ generated as in the description in Section 1.2. We then go on to introduce explicit time indexing and make connections to the PHD filtering recursions of Mahler (2003).

1.3.1 Inference for Partially Observed Poisson Processes

In this subsection we suppress dependence on the parameter θ. Let the realisation of $\mathbf{Y}$ be $\mathbf{y} = \{y_1, \ldots, y_m\}$. In general it is not possible to characterise the distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$, denoted $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$, in closed form; see Lund and Thonnes (2004) for a similar problem solved with perfect simulation. In the case of the Poisson prior, the posterior was characterised only *indirectly* by Mahler (2003) by providing the formula for its Probability Generating Functional (p.g.fl.). Mahler arrived at this formula by differentiating the joint p.g.fl. of the observed and hidden process. Singh et al. (2009) noted that, while this is a general proof technique, it is a technical approach that does not exploit the structure of the problem - a Poisson prior and an observed process constructed via thinning, displacement and augmentation allows for a considerably stronger result with a simpler proof by calling upon several well known results concerning the Poisson PP. Exploiting this Singh et al. (2009) were able to provide a closed-form expression for the posterior which is quite revealing of the "structure"of the conditional process $\mathbf{X}$ given the observed process $\mathbf{Y}$. Corollaries of this result include the expression relating the intensity of the posterior and prior as well as the law of the association of the points of the observed process. While the result in Mahler (2003) is only for a Poisson prior for $\mathbf{X}$, Singh et al. (2009) extends the result to a Gauss-Poisson prior which covers the Poisson prior as a special case. The law $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$ is the foundation of the PHD filter and its derivation presented below follows the approach of Singh et al. (2009).

The derivation of $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$ will draw upon several facts concerning a Poisson PP. The first concerns marking. Let $\mathbf{X}$ be a Poisson PP on E_1 with realisation $\{x_1, \ldots, x_n\}$. Attach to each x_i a random *mark* ζ_i, valued in $\mathcal{M}$ (the mark space) and which is drawn from the probability density $p(\cdot|x_i)$. Additionally, the mark of each point x_i is generated independently. Then $\{(x_1, \zeta_1), \ldots, (x_n, \zeta_n)\}$ is Poisson on $E_1 \times \mathcal{M}$ with intensity $\alpha(x)p(\zeta|x)dxd\zeta$ (Kingman (1993)). Conversely, for a Poisson PP on $E_1 \times \mathcal{M}$ with intensity $v(x, \zeta)$, given the realisation of all the first coordinates, $\{x_1, x_2, \ldots, x_n\}$, then the following is known about the same PP restricted to $\mathcal{M}$. There are n points and they are jointly distributed according to the following density on $\mathcal{M}^n$

$$p(\zeta_1, \ldots, \zeta_n) = \prod_{i=1}^{n} \frac{v(x_i, \zeta_i)}{\int v(x_i, \zeta)d\zeta}. \tag{1.4}$$

According to the multi-target observation model, each point x_i in the realisation of $\mathbf{X}$ generates an E_2-valued observation with probability $p_D(x_i)$. Furthermore, this happens independently for all points in the realisation of $\mathbf{X}$. At this point it is convenient to introduce the following decomposition of $\mathbf{X}$. Two point processes, $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$, are formed from $\mathbf{X}$. $\widehat{\mathbf{X}}$ comprises the points of $\mathbf{X}$ that generate observations

while $\widetilde{\mathbf{X}}$ comprises the remaining unobserved points of $\mathbf{X}$. Since $\widehat{\mathbf{X}}$ is obtained from $\mathbf{X}$ by independent marking then both $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$ are independent Poisson with respective intensities $\alpha(x)p_D(x)$ and $\alpha(x)(1-p_D(x))$ (Kingman, 1993, pp. 55). (The superscript θ on p_D has been omitted from the notation.)

By construction, $\widetilde{\mathbf{X}}$ is unobserved while $\widehat{\mathbf{X}}$ is observed in noise through $\mathbf{Y}$, with noise here referring to the false measurements in $\mathbf{Y}$. This decomposition sheds light on the structure of the posterior: since $\widetilde{\mathbf{X}}$ is unobserved, its law is unchanged after observing $\mathbf{Y}$. As for $\widehat{\mathbf{X}}$, let its posterior be $\mathsf{P}_{\widehat{\mathbf{X}}|\mathbf{y}}$. Thus, the desired posterior $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$ is

$$\mathsf{P}_{\mathbf{X}|\mathbf{y}} = \mathsf{P}_{\widetilde{\mathbf{X}}} * \mathsf{P}_{\widehat{\mathbf{X}}|\mathbf{y}}, \tag{1.5}$$

where $*$ denotes convolution, which follows since $\mathbf{X}$ is the superposition of $\widetilde{\mathbf{X}}$ and $\widehat{\mathbf{X}}$. All that remains to be done is to characterise $\mathsf{P}_{\widehat{\mathbf{X}}|\mathbf{y}}$

Let $\{\Delta\}$ be a one point set with Δ not belonging to either E_1 or E_2 and let $E_1' = E_1 \cup \{\Delta\}$. A marked PP $\mathbf{Z}$ on $E_2 \times E_1'$ is constructed as follows. Each false measurement in $\mathbf{Y}$ is assigned Δ as its mark. Each point in $\mathbf{Y}$ corresponding to a real observation is assigned as a mark the corresponding point in $\widehat{\mathbf{X}}$ that generated it. Let $\mathbf{Z}$ be this set of points formed by marking $\mathbf{Y}$. It follows that $\mathbf{Z}$ is a marked Poisson PP with intensity

$$\alpha(x)p_D(x)g(y|x)\mathbb{I}_{E_2\times E_1}(y,x)dxdy + \kappa(y)\mathbb{I}_{E_2\times\{\Delta\}}(y,x)\delta_\Delta(dx)dy, \tag{1.6}$$

where and $\mathbb{I}_A$ is the indicator function of the set A and $\delta_\Delta(dx)$ is the Dirac measure concentrated on Δ. Given the realisation $\mathbf{y} = \{y_1, \ldots, y_m\}$ of the first coordinate of the process then, by (1.4), the second coordinates are jointly distributed on $(E_1 \cup \{\Delta\})^m$ with law

$$p(dx_1, \ldots, dx_m) = \prod_{i=1}^{m} \frac{\alpha(x_i)p_D(x_i)g(y_i|x_i)\mathbb{I}_{E_1}(x_i)dx_i + \kappa(y_i)\mathbb{I}_{\{\Delta\}}(x)\delta_\Delta(dx_i)}{\int_{E_1}\alpha(x)p_D(x)g(y_i|x)dx + \kappa(y_i)}. \tag{1.7}$$

Theorem 1. *(Singh et al., 2009, Proposition 4.1) Let* $\mathbf{X}$ *be a Poisson PP on* E_1 *with intensity* $\alpha(x)$ *which is observed indirectly through the PP* $\mathbf{Y}$ *on* E_2 *and* $\mathbf{Y}$ *is generated according the observation model detailed in Section 1.2. The conditional distribution of* $\mathbf{X}$ *given the realisation* $\mathbf{y} = \{y_1, \ldots, y_m\}$ *of* $\mathbf{Y}$, $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$, *coincides with the distribution of the superposition of the following two independent point processes:*
– a Poisson PP on E_1 *with intensity* $\alpha(x)(1-p_D(x))dx$ *and*
– the restriction to E_1 *of the an* m*-point PP on* $E_1 \cup \{\Delta\}$ *with law given in (1.7).*

The theorem may be alternately interpreted as follows. To generate a realisation with distribution $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$, the following procedure may be adopted. Generate a realisation of the m-point PP on $E_1 \cup \{\Delta\}$ with law (1.7) by simulating the i-th point according to the i-th measure in the product (1.7). Discard all the points with values Δ and augment this set of remaining points with the realisation of an independent Poisson PP with intensity $\alpha(x)(1-p_D(x))$.

Since $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$ is the law of the superposition of two independent point processes, the following corollary is obvious.

Corollary 1. *(Singh et al., 2009, Proposition 4.1) For a bounded real-valued mea-*

surable function φ on E_1,

$$\begin{aligned}\mathsf{E}\left[\sum_{x\in\mathbf{X}}\varphi(x)\middle|\mathbf{Y}=\mathbf{y}\right] &= \mathsf{E}\left[\sum_{x\in\widetilde{\mathbf{X}}}\varphi(x)\right]+\mathsf{E}\left[\sum_{x\in\widehat{\mathbf{X}}}\varphi(x)\mathbb{I}_{E_1}(x)\middle|\mathbf{Y}=\mathbf{y}\right]\\ &= \int_{E_1}\varphi(x)\alpha(x)(1-p_D(x))dx\\ &\quad+\sum_{i=1}^{m}\frac{\int_{E_1}\varphi(x)\alpha(x)p_D(x)g(y_i|x)dx}{\int_{E_1}\alpha(x)p_D(x)g(y_i|x)dx+\kappa(y_i)}.\end{aligned}$$

When $\varphi(x)=\mathbb{I}_A(x)$ for some subset A of E_1 then the term on the right is precisely the expected number of points in the set A. The non-negative function (on E_1)

$$\alpha(x)(1-p_D(x))+\sum_{i=1}^{m}\frac{\alpha(x)p_D(x)g(y_i|x)}{\int_{E_1}\alpha(x)p_D(x)g(y_i|x)dx+\kappa(y_i)}$$

is the intensity (or first moment) of the PP with law $\mathsf{P}_{\mathbf{X}|\mathbf{y}}$. The intensity of the superposition of two independent processes is the sum of the two intensities and hence the two terms that make up the above expression. This result was first derived, using a different proof technique, in Mahler (2003).

1.3.2 The PHD Filter

The foundations of the PHD filter are the following two facts.

Let $\mathbf{X}_{n-1}$ be a Poisson PP with intensity $\alpha_{n-1}(x_{n-1})$. Since $\mathbf{X}_{n-1}$ evolves to $\mathbf{X}_n$ by a process of independent thinning, displacement and augmentation with an independent Poisson birth process, it follows that marginal distribution of $\mathbf{X}_n$ is also Poisson with intensity (fact 1)

$$\begin{aligned}\alpha_n(x_n) &= \int_{E_1} f^{\theta}(x_n|x_{n-1})p_S^{\theta}(x_{n-1})\alpha_{n-1}(x_{n-1})dx_{n-1}+\gamma^{\theta}(x_n) &(1.8)\\ &=: (\Phi\alpha_{n-1})(x_n)+\gamma^{\theta}(x_n). &(1.9)\end{aligned}$$

This fact may be established using the Thinning, Marking and Superposition theorems for a Poisson process; see Kingman (1993). Specifically, subjecting the realisation of the Poisson PP with intensity α_{n-1} to independent thinning and displacement results in a Poisson PP. The intensity of this PP is given the first function on the right-hand side of (1.9). Combining the realisations of two independent Poisson point processes still results in a Poisson PP. The intensity of the resulting process is the sum of the intensities of the processes being combined. This explains the addition of the term γ^{θ} on the right-hand side of (1.9). Thus, it follows that if $\mathbf{X}_0$ is Poisson then so is $\mathbf{X}_n$ for all n.

It was established in Section 1.3 that the distribution of $\mathbf{X}_1$ conditioned on a realization of observations $\mathbf{y}_1$ is not Poisson. However, the best Poisson approximation to $\mathsf{P}_{\mathbf{X}_1|\mathbf{y}}$, in a Kullback-Leibler sense, is the Poisson PP which has the same intensity as $\mathsf{P}_{\mathbf{X}_1|\mathbf{y}}$ (fact 2); see Mahler (2003), Singh et al. (2009). (This is a general result that applies when any PP is approximated by a Poisson PP using the Kullback-Leibler criterion.) By Corollary 1, the intensity of the best approximating

Poisson PP is

$$\alpha_{1|1}(x_1) = \left[1 - p_D^\theta(x_1) + \sum_{y \in \mathbf{y}_1} \frac{p_D^\theta(x_1) g^\theta(y|x_1)}{\int_{E_1} p_D^\theta(x) g^\theta(y|x) \alpha_{1|0}(x) dx + \kappa^\theta(y)}\right] \alpha_{1|0}(x_1) \tag{1.10}$$

$$=: (\Psi_1 \alpha_{1|0})(x_1) \tag{1.11}$$

where $\alpha_{1|0}$ is the intensity of $\mathbf{X}_1$ and is given by (1.9) with $n = 1$. (Note that no observation is received at time 0.) For convenience in the following we will also write, for each n and $r = 1, 2, ..., m_n$,

$$\mathcal{Z}_{n,r} := \int_{E_n} p_D^\theta(x) g^\theta(y_{n,r}|x) \alpha_{n|n-1}(x) dx.$$

The subscript on the *update* operator Ψ_1 indicates the dependence on the specific realisation of the observations received at time 1. The recursive application of the above two facts gives rise to the PHD filter. Specifically, the conditional distribution of $\mathbf{X}_n$ at each time is approximated by the best fitting Poisson distribution before the subsequent Bayes prediction step. This scheme defines a specific approximation to the optimal filtering recursions for the multi–target model whereby at each time step, only the first moment of the conditional distribution is propagated:

$$\alpha_{n|n-1} = (\Phi \alpha_{n-1|n-1}) + \gamma^\theta, \tag{1.12}$$

$$\alpha_{n|n} = (\Psi_n \alpha_{n|n-1}). \tag{1.13}$$

In the tracking literature these equations are referred to as the PHD filter and were first derived by Mahler (2003). The double subscripts in (1.12), (1.13) imply these are conditional intensities as opposed to the intensity in (1.9), which is the unconditional intensity of the hidden process.

1.4 Approximating the Marginal Likelihood

For a block of realized observations, $\mathbf{y}_{1:n}$, according to the model of section 1.2, we make use of the following decomposition of the marginal likelihood:

$$\begin{aligned} p(\mathbf{y}_{1:n}) &= p(\mathbf{y}_1) \prod_{k=2}^{n} p(\mathbf{y}_k|\mathbf{y}_{1:k-1}) \\ &= \int p(\mathbf{y}_1|\mathbf{x}_1) p(\mathbf{x}_1) d\mathbf{x}_1 \quad \prod_{k=2}^{n} \int p(\mathbf{y}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{y}_{1:k-1}) d\mathbf{x}_k. \end{aligned} \tag{1.14}$$

Using the Poisson approximation of the conditional density $p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ given by the PHD filter, i.e. $\alpha_{k|k-1}$, it follows that the predictive likelihood $p(\mathbf{y}_k|\mathbf{y}_{1:k-1})$ is also Poisson and easily characterised.

Proposition 1. *Let $p(\mathbf{x}_{n-1}|\mathbf{y}_{1:n-1})$ be the density of a Poisson process with intensity $\alpha_{n-1|n-1}$. Then the predictive density of the observation $\mathbf{y}_n$,*

$$p(\mathbf{y}_n|\mathbf{y}_{1:n-1}) = \int p(\mathbf{y}_n|\mathbf{x}_n) p(\mathbf{x}_n|\mathbf{y}_{1:n-1}) d\mathbf{x}_n,$$

is the density of a Poisson process with intensity function given by:

$$\int_{E_1} g^\theta(y|x_n)p_D^\theta(x_n)\alpha_{n|n-1}(x_n)dx_n + \kappa^\theta(y). \tag{1.15}$$

Proof. Recall from the definition of the model that, given $\mathbf{x}_{n-1}$, $\mathbf{X}_n$ is formed as follows. With probability $p_S^\theta(x)$, each point of $\mathbf{x}_{n-1}$ survives and mutates according the Markov kernel $f^\theta(\cdot|x_{n-1})$. This happens independently for each point of $\mathbf{x}_{n-1}$. $\mathbf{X}_n$ then consists of the surviving and mutated points of $\mathbf{x}_{n-1}$, superposed with points from an independent birth Poisson process with intensity γ^θ. From the Thinning, Marking and Superposition theorems for a Poisson processes (see Kingman (1993)), and under the condition of the proposition, $p(\mathbf{x}_n|\mathbf{y}_{1:n-1})$ is then Poisson with intensity $\alpha_{n|n-1}$ as defined in (1.12). The observation $\mathbf{Y}_n$ is then formed as follows. With probability $p_D^\theta(x)$, each point of $\mathbf{X}_n$ is detected and generates an observation in E_2 through the probability density $g^\theta(\cdot|x_n)$. $\mathbf{Y}_n$ then consists of the observations originating from $\mathbf{X}_n$, superposed with an independent clutter Poisson process of intensity κ^θ. It follows once again from the Thinning, Marking and Superposition theorems that, under the condition of the proposition, $p(\mathbf{y}_n|\mathbf{y}_{1:n-1})$ is Poisson with intensity given by (1.15). □

For a realized Poisson process $\mathbf{y} = y_{1:k}$ in E_2 with intensity function $\beta(y)$, the likelihood is given by:

$$p(\mathbf{y}) = \frac{1}{k!}\exp\left[-\int \beta(y)dy\right]\prod_{j=1}^{k}\beta(y_j). \tag{1.16}$$

Combining (1.14), Proposition 1 and (1.16), the log-likelihood of the observed data may be approximated as follows:

$$\begin{aligned}\ell_{\mathrm{Po},n}(\theta) = &-\sum_{k=1}^{n}\int\left[\int g^\theta(y_k|x_k)p_D^\theta(x_k)\alpha_{k|k-1}(x_k)dx_k + \kappa^\theta(y_k)\right]dy_k \\ &+\sum_{k=1}^{n}\sum_{r=1}^{m_k}\log\left[\int g^\theta(y_{k,r}|x_k)p_D^\theta(x_k)\alpha_{k|k-1}(x_k)dx_k + \kappa^\theta(y_{k,r})\right],\end{aligned} \tag{1.17}$$

where the subscript Po on ℓ indicates that this is an approximation based on the Poisson approximations to $p(\mathbf{x}_n|\mathbf{y}_{1:n-1})$.

1.5 SMC approximation of the PHD filter and its gradient

In the majority of cases, the Poisson approximation (1.17) of the true log-likelihood $\log p(\mathbf{y}_{1:n})$ cannot be computed exactly and a numerical scheme to approximate the various integrals therein is needed. It was noted by Vo and Ma (2006) that under certain conditions on the multi-target model, the predicted and updated intensities are Gaussian mixtures and the recursion (1.12)–(1.13) is analytically tractable. However, the number of components in these mixtures explodes over time and so Vo and Ma (2006) employed a pruning mechanism to allow practical implementation. Extended Kalman Filter and the Unscented Kalman Filter style deterministic approximations of the intensity recursions have also been devised to cope with a more general model. Predating these works is Vo et al. (2005) where a SMC method to approximate the intensity functions was devised. In this section

we review this original SMC algorithm and extend it to approximate the gradient of the intensity functions as well. The SMC approximation of the PMT likelihood and its gradient is also detailed.

SMC methods have become a standard tool for computation in non-linear optimal filtering problems and in this context have been termed particle filters. We do not give explicit details of standard particle filtering algorithms here but refer the reader to Doucet et al. (2001) and Cappé et al. (2007) for a variety of algorithms, theoretical details and applications, and Del Moral et al. (2006) for a general framework. SMC algorithms may be viewed as being constructed from ideas of Sequential Importance Sampling (SIS) and resampling. They recursively propagate a set of weighted random samples called particles, which are used to approximate a sequence of probability distributions. The algorithms are such that, as the number of particles tends to infinity and under weak assumptions, an integral with respect to the random distribution defined by the particle set converges to the integral with respect to the corresponding true distribution.

A particle implementation of the PHD filter (equations (1.12)-(1.13)) was proposed simultaneously in several works; Vo et al. (2003), Siddenblath (2003), Zajic and Mahler (2003). These implementations may be likened to the Bootstrap Particle filter in the sense that their "proposal" steps ignore the new observations. An auxiliary SMC implementation that takes into account the new observations at each time was recently proposed by Whiteley, Singh, and Godsill (Whiteley et al.) to minimise the variance of the incremental weight. The particle algorithm of Vo et al. (2005) will be the building block for the particle approximation of (1.17) and its gradient.

Given the particle set from the previous iteration, $\{X_{n-1}^{(i)}, W_{n-1}^{(i)}\}_{i=1}^{N}$, N samples, $\{X_n^{(i)}\}_{i=1}^{N}$ are each drawn from a proposal distribution $q_n(\cdot|X_{n-1}^{(i)})$ and the predicted importance weights $\{W_{n|n-1}^{(i)}\}_{i=1}^{N}$ are computed as follows:

$$X_n^{(i)} \sim q_n(\cdot|X_{n-1}^{(i)}), \qquad W_{n|n-1}^{(i)} = \frac{f^{\theta}(X_n^{(i)}|X_{n-1}^{(i)})p_S^{\theta}(X_{n-1}^{(i)})}{q_n(X_n^{(i)}|X_{n-1}^{(i)})} W_{n-1}^{(i)}, \qquad 1 \leq i \leq N. \tag{1.18}$$

This proposal distribution can depend on θ, e.g. set q_n to be the transition density for the individual targets f^{θ}, which was implemented in the numerical examples in Section 1.7. A collection of L additional samples, $\{X_n^{(i)}\}_{i=N+1}^{N+L}$, dedicated to the birth term are then drawn from a proposal distribution $p_n(\cdot)$ and the corresponding importance weights $\{W_{n|n-1}^{(i)}\}_{i=N+1}^{N+L}$ are computed:

$$X_n^{(i)} \sim p_n(\cdot), \qquad W_{n|n-1}^{(i)} = \frac{1}{L}\frac{\gamma^{\theta}(X_n^{(i)})}{p_n(X_n^{(i)})}, \qquad N+1 \leq i \leq N+L. \tag{1.19}$$

p_n can also depend on θ and the default choice would be proportional to γ^{θ} (provided the corresponding normalised density exists and can be sampled from easily). The approximation to the predicted intensity $\alpha_{n|n-1}$ is

$$\widehat{\alpha}_{n|n-1}(dx_n) = \sum_{i=1}^{N+L} W_{n|n-1}^{(i)} \delta_{X_n^{(i)}}(dx_n).$$

$\widehat{\alpha}_{n|n-1}$ may be used to approximate integrals of the form $\int_E \psi(x)\alpha_{n|n-1}(x)dx$ which appear in the denominator of (1.13). For $r \in \{1, 2, ..., m_n\}$, this approximation is given by:

$$\widehat{\mathcal{Z}}_{n,r} = \sum_{i=1}^{N+L} \psi_{n,r}(X_n^{(i)}) W_{n|n-1}^{(i)} + \kappa^{\theta}(y_{n,r}), \tag{1.20}$$

where

$$\psi_{n,r}(x) = p_D^{\theta}(x) g^{\theta}(y_{n,r}|x). \tag{1.21}$$

The particles are then re-weighted according to the update operator yielding a second collection of importance weights $\{W_n^{(i)}\}_{i=1}^{N+L}$ defined as follows:

$$W_n^{(i)} = \left[1 - p_D^{\theta}(X_n^{(i)}) + \sum_{r=1}^{m_n} \frac{\psi_{n,r}(X_n^{(i)})}{\widehat{\mathcal{Z}}_{n,r}}\right] W_{n|n-1}^{(i)}. \tag{1.22}$$

The empirical measure defined by the particle set $\{X_n^{(i)}, W_n^{(i)}\}_{i=1}^{N+L}$ then approximates the updated intensity $\alpha_{n|n}$:

$$\widehat{\alpha}_{n|n}(dx_n) := \sum_{i=1}^{N+L} W_n^{(i)} \delta_{X_n^{(i)}}(dx_n). \tag{1.23}$$

The importance weights, with total mass $\sum_{i=1}^{N+L} W_n^{(i)}$, are then normalised so that they sum to 1, and after resampling N times to obtain $\{X_n'^{(i)}\}_{i=1}^{N}$, the importance weights are set to the constant $(\sum_{i=1}^{N+L} W_n^{(i)})/N$. Vo et al. (2005) also noted that the total number of particles may be varied across iterations, perhaps guided by the total mass of the updated intensity. Convergence results establishing the theoretical validity of the particle PHD filter have been obtained. Convergence of expected error was established in Vo et al. (2005), almost sure convergence and convergence of mean-square error were established in Clark and Bell (2006) and $\mathbb{L}_p$ error bounds, almost sure convergence and a Central Limit Theorem were established in Johansen et al. (2006).

Because of the low dimension of $\alpha_{n|n}$ (e.g. four when the state descriptor of individual targets contains position and velocity only) even a simple SMC implementation like the one outlined above may suffice. In the case when the observations are informative, and the likelihood functions are concentrated, weight degeneracy can occur with the above implementation and the auxiliary version of Whiteley, Singh, and Godsill (Whiteley et al.) has been shown to be more efficient.

For the gradient of the PHD filter, we will use the method proposed in Poyiadjis et al. (2005), Poyiadjis et al. (2009) for the closely related problem of computing the gradient of the log-likelihood for a HMM. Let $\overline{\nabla(\Phi\alpha_{n-1|n-1})}$ be a *pointwise* approximation of $\nabla(\Phi\alpha_{n-1|n-1})$, that is $\nabla(\Phi\alpha_{n-1|n-1})(x_n) \approx \overline{\nabla(\Phi\alpha_{n-1|n-1})}(x_n)$, $x_n \in E_1$. (This pointwise approximation is constructed in a sequential manner as detailed below.) One possible construction of a particle approximation to $\nabla\alpha_{n|n-1}(x_n)dx_n$ is

the following:

$$\widehat{\nabla\alpha}_{n|n-1}(dx_n) = \frac{1}{N}\sum_{i=1}^{N}\frac{\overline{\nabla(\Phi\alpha_{n-1|n-1})}(X_n^{(i)})}{Q_n(X_n^{(i)})}\delta_{X_n^{(i)}}(x_n) + \frac{1}{L}\sum_{i=N+1}^{N+L}\frac{\nabla\gamma^\theta(X_n^{(i)})}{p_n(X_n^{(i)})}\delta_{X_n^{(i)}}(x_n) \tag{1.24}$$

where

$$Q_n(x_n) = \frac{1}{\sum_{j=1}^{N+L}W_{n-1}^{(j)}}\sum_{i=1}^{N+L}q_n(x_n|X_{n-1}^{(i)})W_{n-1}^{(i)}.$$

Note that the particle set $\{X_n^{(i)}\}_{i=1}^N$ in (1.18) was obtained by sampling Q_n N times. (Assuming (1.23) is resampled at every time n.) Re-write the particle approximation to $\nabla\alpha_{n|n-1}$ as

$$\widehat{\nabla\alpha}_{n|n-1} = \sum_{i=1}^{N+L}\delta_{X_n^{(i)}}(x_n)A_{n|n-1}^{(i)}W_{n|n-1}^{(i)}$$

where

$$A_{n|n-1}^{(i)} = \frac{1}{N}\frac{\overline{\nabla(\Phi\alpha_{n-1|n-1})}(X_n^{(i)})}{Q_n(X_n^{(i)})}\frac{1}{W_{n|n-1}^{(i)}}, \qquad 1\le i\le N,$$

$$A_{n|n-1}^{(i)} = \frac{\nabla\gamma^\theta(X_n^{(i)})}{\gamma^\theta(X_n^{(i)})}, \qquad N+1\le i\le N+L.$$

The pointwise approximation to $\nabla(\Phi\alpha_{n-1|n-1})$ is

$$\begin{aligned}&\overline{\nabla(\Phi\alpha_{n-1|n-1})}(x_n)\\ &= \int_{E_1}\left[\nabla\log f^\theta(x_n|x_{n-1})\right]f^\theta(x_n|x_{n-1})p_S^\theta(x_{n-1})\widehat{\alpha}_{n-1|n-1}(dx_{n-1})\\ &= \int_{E_1}\left[\nabla\log p_S^\theta(x_{n-1})\right]f^\theta(x_n|x_{n-1})p_S^\theta(x_{n-1})\widehat{\alpha}_{n-1|n-1}(dx_{n-1})\\ &+ \int_{E_1}f^\theta(x_n|x_{n-1})p_S^\theta(x_{n-1})\widehat{\nabla\alpha}_{n-1|n-1}(dx_{n-1}).\end{aligned} \tag{1.25}$$

Using (1.24), the particle approximation to $\nabla\int\psi_{n,r}(x_n)\alpha_{n|n-1}(x_n)dx_n + \nabla\kappa^\theta(y_r)$, for $r = 1, 2, \ldots, m_n$, is

$$\begin{aligned}\widehat{\nabla\mathcal{Z}}_{n,r} &= \int\nabla\psi_{n,r}(x_n)\widehat{\alpha}_{n|n-1}(dx_n)\\ &+ \int\psi_{n,r}(x_n)\widehat{\nabla\alpha}_{n|n-1}(dx_n) + \nabla\kappa^\theta(y_r)\\ &= \sum_{i=1}^{N+L}\left(\nabla\psi_{n,r}(X_n^{(i)}) + \psi_{n,r}(X_n^{(i)})A_{n|n-1}^{(i)}\right)W_{n|n-1}^{(i)} + \nabla\kappa^\theta(y_r).\end{aligned}$$

The particle approximation of $\nabla\alpha_{n|n}$ is constructed by re-weighting the particle approximations of $\nabla\alpha_{n|n-1}$ and $\alpha_{n|n-1}$:

$$\begin{aligned}\widehat{\nabla\alpha}_{n|n}(dx_n) &= -\nabla p_D^\theta(x_n)\widehat{\alpha}_{n|n-1}(dx_n) \\ &+ \left[\sum_{r=1}^{m_n}\frac{\psi_{n,r}(x_n)}{\widehat{\mathcal{Z}}_{n,r}}\left(\nabla\log\psi_{n,r}(x_n) - \frac{\widehat{\nabla\mathcal{Z}}_{n,r}}{\widehat{\mathcal{Z}}_{n,r}}\right)\right]\widehat{\alpha}_{n|n-1}(dx_n) \\ &+ \left[1 - p_D^\theta(x_n) + \sum_{r=1}^{m_n}\frac{\psi_{n,r}(x_n)}{\widehat{\mathcal{Z}}_{n,r}}\right]\widehat{\nabla\alpha}_{n|n-1}(dx_n) \\ &= \sum_{i=1}^{N+L} A_n^{(i)} W_n^{(i)}\delta_{X_n^{(i)}}(dx_n) \end{aligned} \tag{1.26}$$

where

$$\begin{aligned}A_n^{(i)} &= A_{n|n-1}^{(i)} \\ &+ \left[-\nabla p_D^\theta(X_n^{(i)}) + \sum_{r=1}^{m_n}\frac{\psi_{n,r}(X_n^{(i)})}{\widehat{\mathcal{Z}}_{n,r}}\left(\nabla\log\psi_{n,r}(X_n^{(i)}) - \frac{\widehat{\nabla\mathcal{Z}}_{n,r}}{\widehat{\mathcal{Z}}_{n,r}}\right)\right] \\ &\times \frac{W_{n|n-1}^{(i)}}{W_n^{(i)}}.\end{aligned}$$

The SMC estimate of ℓ_{Po} (for the same θ used in the weight calculation above and proposal distributions above) is given by:

$$\widehat{\ell}_{\mathrm{Po},n}(\theta) = -\sum_{k=1}^{n}\left[\int p_D^\theta(x_k)\widehat{\alpha}_{k|k-1}(dx_k) + \int\kappa^\theta(y_k)dy_k\right] + \sum_{k=1}^{n}\sum_{r=1}^{m_k}\log\widehat{\mathcal{Z}}_{k,r} \tag{1.27}$$

And the estimate of $\nabla\ell_{\mathrm{Po},n}$ is

$$\begin{aligned}&\widehat{\nabla\ell}_{\mathrm{Po},n}(\theta) \\ &= -\sum_{k=1}^{n}\int\left[\sum_{i=1}^{N+L}\left(\nabla p_D^\theta(X_k^{(i)}) + p_D^\theta(X_k^{(i)})A_{k|k-1}^{(i)}\right)W_{k|k-1}^{(i)} + \int\nabla\kappa^\theta(y_k)dy_k\right] \\ &+ \sum_{k=1}^{n}\sum_{r=1}^{m_k}\frac{\widehat{\nabla\mathcal{Z}}_{k,r}}{\widehat{\mathcal{Z}}_{k,r}}.\end{aligned}$$

Algorithm 1 summarises the proposed SMC method for computing $\widehat{\ell}_{\mathrm{Po},n}$ and $\widehat{\nabla\ell}_{\mathrm{Po},n}$. The computational cost of this algorithm, unlike a conventional particle filter, grows quadratically in the number of particles N.

1.6 Parameter Estimation

1.6.1 Pointwise Gradient Approximation

Equipped with an approximation of the true likelihood $p(\mathbf{y}_{1:n})$ and its gradient, the parameters of the model may be estimated with a gradient ascent algorithm. This may be done in an off-line fashion once a batch of observations, say $\mathbf{y}_{1:T}$, has been received, or in an online manner. This section discusses both these methods of estimation.

Let the true static parameter generating the sequence of observations be θ^* and it is to be estimated from the observed data $\{\mathbf{y}_n\}_{n\geq 1}$. Given the a record

Algorithm 1 Particle Approximation of the Intensity and its Sensitivity

At time 0

for $i = 1$ to N **do**

$X_0^{(i)} \sim q_0(\cdot)$

$W_0^{(i)} = \frac{1}{N}$

$A_0^{(i)} = 0$

end for

Set $\ell_{\mathrm{Po},0} = 0$, $\widehat{\nabla \ell}_{\mathrm{Po}} = [0, \ldots, 0]$ $(\in \mathbf{R}^d)$

At time $n \geq 1$

Prediction Step:

for $i = 1$ to N **do**

$X_n^{(i)} \sim q_n(\cdot | X_{n-1}^{(i)})$

Set $W_{n|n-1}^{(i)} = \dfrac{f^\theta(X_n^{(i)}|X_{n-1}^{(i)}) p_S^\theta(X_{n-1}^{(i)})}{q_n(X_n^{(i)}|X_{n-1}^{(i)})} W_{n-1|n-1}^{(i)}$

Set $A_{n|n-1}^{(i)} W_{n|n-1}^{(i)} = \dfrac{1}{N} \dfrac{\overline{\nabla(\Phi \alpha_{n-1|n-1})}(X_n^{(i)})}{Q_n(X_n^{(i)})}$

end for

for $i = N+1$ to $N+L$ **do**

$X_n^{(i)} \sim p_n(\cdot)$

Set $W_{n|n-1}^{(i)} = \dfrac{1}{L} \dfrac{\gamma^\theta(X_n^{(i)})}{p_n(X_n^{(i)})}$, $\quad A_{n|n-1}^{(i)} W_{n|n-1}^{(i)} = \dfrac{\nabla \gamma^\theta(X_n^{(i)})}{p_n(X_n^{(i)})}$

end for

for $r = 1$ to m_n **do**

Set $\widehat{\mathcal{Z}}_{n,r} = \sum_{i=1}^{N+L} \psi_{n,r}(X_n^{(i)}) W_{n|n-1}^{(i)} + \kappa^\theta(y_{n,r})$

Set $\widehat{\nabla \mathcal{Z}}_{n,r} = \sum_{i=1}^{N+L} \left(\nabla \psi_{n,r}(X_n^{(i)}) + \psi_{n,r}(X_n^{(i)}) A_{n|n-1}^{(i)} \right) W_{n|n-1}^{(i)} + \nabla \kappa^\theta(y_r)$

end for

Weight Step:

for $i = 1$ to $N+L$ **do**

Set $W_n^{(i)} = \left[1 - p_D^\theta(X_n^{(i)}) + \sum_{r=1}^{m_n} \dfrac{\psi_{n,r}(X_n^{(i)})}{\widehat{\mathcal{Z}_{n,r}}} \right] W_{n|n-1}^{(i)}$

Set $A_n^{(i)} = A_{n|n-1}^{(i)} +$

$\left[-\nabla p_D^\theta(X_n^{(i)}) + \sum_{r=1}^{m_n} \dfrac{\psi_{n,r}(X_n^{(i)})}{\widehat{\mathcal{Z}}_{n,r}} \left(\nabla \log \psi_{n,r}(X_n^{(i)}) - \dfrac{\widehat{\nabla \mathcal{Z}_{n,r}}}{\widehat{\mathcal{Z}}_{n,r}} \right) \right] \dfrac{W_{n|n-1}^{(i)}}{W_n^{(i)}}$

end for

Resample $\left\{ X_n^{(i)}, \dfrac{W_n^{(i)}}{\sum_{j=1}^{N+L} W_n^{(j)}} \right\}_{i=1}^{N+L}$ to obtain $\left\{ X_n^{(i)}, \dfrac{\sum_{j=1}^{N+L} W_n^{(j)}}{N} \right\}_{i=1}^{N}$

Compute Likelihood and Likelihood Gradient:

$\widehat{\ell}_{\mathrm{Po},n} = \widehat{\ell}_{\mathrm{Po},n-1} - \int \sum_{i=1}^{N+L} p_D^\theta(X_n^{(i)}) W_{n|n-1}^{(i)} - \int \kappa^\theta(y_n) dy_n + \sum_{r=1}^{m_n} \log \widehat{\mathcal{Z}_{n,r}}$

$\widehat{\nabla \ell}_{\mathrm{Po},n} = \widehat{\nabla \ell}_{\mathrm{Po},n-1} - \sum_{i=1}^{N+L} \left(\nabla p_D^\theta(X_n^{(i)}) + p_D^\theta(X_n^{(i)}) A_{n|n-1}^{(i)} \right) W_{n|n-1}^{(i)} - \int \nabla \kappa^\theta(y_n) dy_n + \sum_{r=1}^{m_n} \dfrac{\widehat{\nabla \mathcal{Z}}_{n,r}}{\widehat{\mathcal{Z}}_{n,r}}$.

of observations $\{\mathbf{y}_n\}_{n=1}^T$, the log-likelihood may be maximized with the following steepest ascent algorithm. For a discrete time index, $k = 1, 2, ...$, which does not coincide with the time index of the observation sequence,

$$\theta_{k+1} = \theta_k + a_{k+1} \nabla \ell_{\mathrm{Po},T}(\theta)|_{\theta=\theta_k}, \qquad k \geq 1, \tag{1.28}$$

where $\{a_k\}_{k\geq 1}$ is a sequence of small positive real numbers, called the step-size sequence, that should satisfy the following constraints: $\sum_k a_k = \infty$ and $\sum_k a_k^2 < \infty$. One possible choice would be $a_k = k^{-\zeta}$, $0.5 < \zeta < 1$ (e.g. $a_k = k^{-2/3}$); see Pflug (1996) for background theory on steepest ascent.

For a long observation sequence the computation of the gradient in (1.28) can be prohibitively expensive. A more attractive alternative would be a recursive procedure in which the data is run through once sequentially. For example, consider the following update scheme:

$$\theta_{n+1} = \theta_n + a_{n+1} \nabla \log p_{\mathrm{Po}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})|_{\theta=\theta_n}. \tag{1.29}$$

Upon receiving y_n, θ_n is updated in the direction of ascent of the conditional density of this new observation. The algorithm in the present form is not suitable for online implementation due to the need to evaluate the gradient of $\log p_{\mathrm{Po}}(y_n|y_{1:n-1})$ at the current parameter estimate. Doing so would require browsing through the entire history of observations. This limitation is removed by computing $\nabla \log p_{\mathrm{Po}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})|_{\theta=\theta_n}$ recursively using the previous values of the parameter as well. This modification is straightforward; see Poyiadjis et al. (2009) for the closely related problem of recursive maximum likelihood estimation in HMMs.

In practice, it may be beneficial to start with a constant but small step-size, $a_n = a$ for some initial period $n < n^*$. If the step-size decreases too quickly the algorithm might get stuck at an early stage and fail to come close to a global maximum of the likelihood.

1.6.2 Simultaneous Perturbation Stochastic Approximation (SPSA)

It is also possible to maximise ℓ_{Po} without explicit computation of the gradient using SMC. In particular, a finite difference (FD) approximation of the gradient may be constructed from the noisy evaluations of ℓ_{Po} obtained using SMC. Such approaches are often termed "gradient-free"; see Spall (2003).

Consider the problem of maximizing a real valued function $\theta \in \Theta \to \ell(\theta)$ where Θ is an open subset of $\mathbb{R}$. The first steepest ascent algorithm based on FD approximation of the likelihood gradient is due to Kiefer and Wolfowitz (1952). This involves, for example in the two-sided case, noisy evaluation of ℓ at perturbed parameter values $\theta \pm \Delta$ and the subsequent approximation of the gradient at this parameter value, $\nabla\ell(\theta)$, as follows:

$$\widehat{\nabla\ell}(\theta) = \frac{\widehat{\ell}(\theta+\Delta) - \widehat{\ell}(\theta-\Delta)}{2\Delta},$$

The method can be generalized to the case in which $\Theta \subset \mathbb{R}^d$, $d \geq 1$, by carrying out two noisy evaluations of ℓ for each dimension of Θ, but this can become computationally expensive when the dimension is high. An alternative is the SPSA method of Spall (1992), which requires only two noisy evaluations of ℓ, regardless of the dimension of Θ. This method takes its name from the fact that it involves perturbing the multiple components of the vector θ at the same time. In the case of SPSA, the estimate of the gradient at the k-th iteration of the gradient ascent algorithm (i.e. the recursion (1.28)) is:

$$\widehat{\nabla_p \ell}_{\mathrm{Po},T}(\theta_k) = \frac{\widehat{\ell}_{\mathrm{Po},T}(\theta_k + c_k\Delta_k) - \widehat{\ell}_{\mathrm{Po},T}(\theta_k - c_k\Delta_k)}{2c_k\Delta_{k,p}}, \qquad p = 1, \ldots, d,$$

where $\Delta_k = [\Delta_{k,1}\ \Delta_{k,2}\ \ldots\ \Delta_{k,d}]^T$ is a random perturbation vector, $\{c_k\}_{k\geq 0}$ is a decreasing sequence of small positive numbers and $\nabla_p \ell_{\mathrm{Po},T}$ is the partial derivative of $\ell_{\mathrm{Po},T}$ w.r.t. the p-th component of the parameter θ. The elements of Δ_k are i.i.d., non-zero, symmetrically distributed random variables. In this case we take them to be Bernoulli ± 1 distributed, but it should be noted that some alternative choices, such as zero-mean Gaussian distributed are theoretically invalid, see (Spall, 1992, Chapter 7) for further background details. The objective function $\ell_{\mathrm{Po},T}(\theta)$ is estimated using the SMC implementation of Section 5. Theoretical results guaranteeing convergence of SPSA require the following conditions on the gain sequences (Spall (2003)):

$$\forall\ k,\ a_k > 0 \text{ and } c_k > 0; \quad a_k \text{ and } c_k \to 0; \quad \sum_{k=0}^{\infty} a_k = \infty; \quad \sum_{k=0}^{\infty} \frac{a_k^2}{c_k^2} < \infty. \tag{1.30}$$

Practical choices of these sequences can be based around the following expressions, advocated and related to theoretical properties of SPSA in Spall (2003). For non-negative coefficients a, c, A, ς, τ:

$$a_k = \frac{a}{(k+1+A)^{\varsigma}}, \qquad c_k = \frac{c}{(k+1)^{\tau}}.$$

The recommendation is to set $\varsigma = 0.6$ and $\tau = 0.1$ and, as a rule of thumb, to choose A to be 10% or less of the maximum number of allowed iterations of the steepest ascent recursion.

Throughout the simulation study in section 1.7.3, common random numbers were used for each pair of noisy evaluations of the objective function. It has been shown by Kleinman et al. (1999) that using common random numbers in this way leads to faster convergence of the steepest ascent algorithm. It should be noted that a number of other strategies, such as iterate averaging and adaptive schemes involving estimation of the Hessian matrix, can improve the performance of SPSA. These techniques are beyond the scope of this chapter and are discussed in Spall (2003).

SPSA for maximising $\ell_{\mathrm{Po},T}$ is summarized in Algorithm 2 below.

Algorithm 2 SPSA Parameter Estimation

At time 1
  Initialize θ_1
At time $k \geq 1$
  Generate Perturbation vector Δ_k
  Run SMC PHD Filter with $\theta = \theta_k + c_k\Delta_k$
    Compute $\widehat{\ell}_{\mathrm{Po},T}(\theta_k + c_k\Delta_k)$ according to (1.27)
  Run SMC PHD Filter with $\theta = \theta_k - c_k\Delta_k$
    Compute $\widehat{\ell}_{\mathrm{Po},T}(\theta_k - c_k\Delta_k)$ according to (1.27)
  Set $\widehat{\nabla}\ell_{\mathrm{Po},T}(\theta_k) = \frac{\widehat{\ell}_{\mathrm{Po},T}(\theta_k+c_k\Delta_k) - \widehat{\ell}_{\mathrm{Po},T}(\theta_k-c_k\Delta_k)}{2c_k}[\frac{1}{\Delta_{k,1}}\ \frac{1}{\Delta_{k,2}}\ \cdots\ \frac{1}{\Delta_{k,d}}]^T$
  Set $\theta_{k+1} = \theta_k + a_k\widehat{\nabla}\ell_{\mathrm{Po},T}(\theta_k)$

1.7 Simulation Study

1.7.1 Model

The proposed parameter estimation methods are evaluated on a multi-target model with the following characteristics.

A constant velocity model is assumed for individual targets. The position of a target is specified in two dimensions, restricted to the window $[0, 100] \times [0, 100]$. The state of a single target is thus specified by a 4 dimensional vector $X_n = [X_n(1), X_n(2), X_n(3), X_n(4)]^T$; the variables $(X_n(1), X_n(3))$ specify position and $(X_n(2), X_n(4))$ specify velocity. The state of the single target evolves over time as follows:

$$X_n = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} X_{n-1} + \begin{bmatrix} V_n(1) \\ V_n(2) \\ V_n(3) \\ V_n(4) \end{bmatrix}$$

where $V_n(1)$ and $V_n(3)$ are independent Gaussian random variables with mean 0 and standard deviation $\sigma_{xs} = 0.01$. $V_n(2)$ and $V_n(4)$ are also independent Gaussian random variables with mean 0 and standard deviation $\sigma_{xv} = 0.25$. The state of individual targets is a vector in $E_1 = [0, 100] \times \mathbb{R} \times [0, 100] \times \mathbb{R}$. The birth intensity is defined as

$$\gamma^{\theta}(\cdot) = \Gamma \mathcal{N}(\cdot; \mu_b, \Sigma_b), \tag{1.31}$$

where $\mathcal{N}(x; \mu_b, \Sigma_b)$ denotes the multivariate normal density with mean μ_b and covariance Σ_b, evaluated at x. For the numerical example,

$$\mu_b = \begin{bmatrix} \mu_{bx} \\ 0 \\ \mu_{by} \\ 0 \end{bmatrix}, \quad \Sigma_b = \begin{bmatrix} \sigma_{bs}^2 & 0 & 0 & 0 \\ 0 & \sigma_{bv}^2 & 0 & 0 \\ 0 & 0 & \sigma_{bs}^2 & 0 \\ 0 & 0 & 0 & \sigma_{bv}^2 \end{bmatrix}. \tag{1.32}$$

The x–y position of the target is observed in additive, isotropic Gaussian noise with standard deviation σ_y. The clutter intensity is $\kappa(y)$ is uniform on $[0, 100] \times [0, 100]$:

$$\kappa(y) = \kappa \mathbb{I}_{[0,100]\times[0,100]}(y). \tag{1.33}$$

The probability of detection $p_D(x)$ and survival $p_S(x)$ is assumed constant over E_1. The measurements from individual targets and clutter are vectors in $E_2 = \mathbb{R}^2$.The parameters of the model to be inferred from the data in the numerical studies below are

$$\theta = [\sigma_y,\ \kappa,\ \Gamma,\ \mu_{bx},\ \mu_{by},\ \sigma_{bs},\ \sigma_{bv},\ p_D]^T$$

with $(\sigma_{xs}, \sigma_{xv}, p_S)$ assumed known.

1.7.2 Pointwise Gradient Approximation

The performance of gradient ascent, (see (1.28)) with Algorithm 1 to estimate the derivative of the likelihood, was evaluated using the following set of parameters: $\theta^* = [5, 4 \times 10^{-4}, 1, 50, 50, 5, 2, 0.9]^T$, i.e. the observation record was generated

using these parameter values. For an observation time of length 50, these values for the model would generate, on average, 9 observations per time instant with 4 of them being false, and a total of 50 targets for all the 50 time points. The number of particles used were $N = 400$ and $L = 400$.

Figure 1.1 shows the sequence of iterates generated by (1.28) for a constant step-size sequence, i.e. $a_k = a$, and for an initial value chosen to be reasonably distant from θ^* (consult the figure for the initial values). The estimated model parameters converge to a vicinity of the true parameters but with some notable discrepancies. (Further details in the discussion to follow.) The smoothness of the traces also indicate the that estimate of the gradient with Algorithm 1 has low variance.

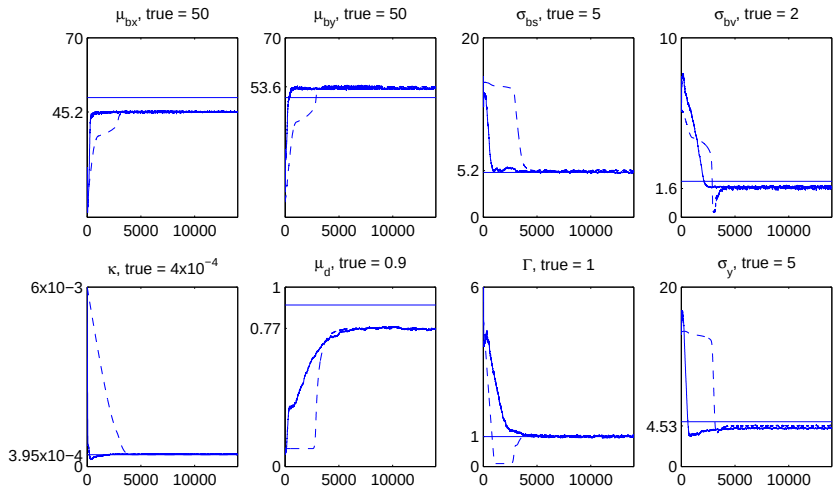

Figure 1.1: Evolution of the parameter estimates generated by steepest ascent with the gradient computed using Algorithm 1 (dashed) and SPSA (solid). Observation record length was 25 generated with the model outlined at the start of Section 1.7. Only the values after every 50–th step are displayed. True value of parameters are indicated by the horizontal lines and the estimated values of the gradient based method are marked on the y-axis. Notable discrepancies for μ_{bx} and μ_{by} would be outliers for a longer observation record; see box plots in Figure 1.2.

To characterise the distribution of the estimated parameters, the experiment was repeated a total of 50 times for observation records of length 15, 50 and 125. In each repetition of the experiment, the targets and observation record were generated again from the model. The distribution of the converged value for the parameters are shown in Figure 1.2 when the gradient is approximated using Algorithm 1 along with their true values as horizontal lines. As can be seen from the box plots, the estimated model parameters do improve with longer observation records. The results are encouraging and are a further verification of the merit of the Poisson approximation of the posterior in Proposition 1; thus far the approximation has only been verified by comparing the estimated intensity with the true number of targets in synthetic numerical examples. It is evident from the box plots in Figure 1.2 that there are small biases for some of the parameters. It is unclear if these are due to the insufficient number of particles used to approximate the intensity function or is inherent to the approximate likelihood itself (see (1.17)). For example, the box plots indicate a bias in the estimation of the clutter intensity; it is being over estimated. This may be explained by the small number of particles used in the simulation. On average, for an observation of length 50, there were 9 targets at any particular time instant. 400 particles are not sufficient to follow the all the modes of the intensity

function (1.13) induced by these targets. Hence the observations generated by the targets are perhaps being accounted for by a higher clutter intensity estimate. We also remark that the converged values of the estimates for μ_{bx} and in μ_{by} in Figure 1.1 would be outliers for an observation record of length 50. This can be seen from the corresponding box plots in Figure 1.2.

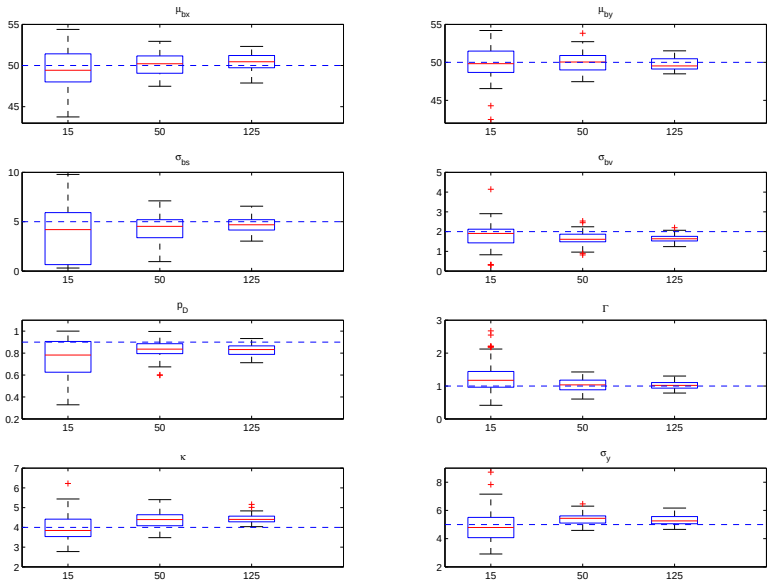

Figure 1.2: Box plots of the converged value of parameter estimates for different observation record lengths. Dashed horizontal lines indicate the true value of the parameters.

1.7.3 SPSA

The SPSA scheme, algorithm 2, with algorithm 1 used to obtain a noisy evaluation of the Poisson likelihood, was run on the same data record as the pointwise gradient method, with the same initial conditions. A constant step size of $a_k = a = 1 \times 10^{-5}$ and $c_k = c = 0.025$ were chosen after a few pilot runs.

A fixed step size was chosen so as to avoid premature convergence of the algorithm. After 20,000 iterations, and having reached equilibrium, the SPSA scheme resulted in the following parameter values: $\mu_{bx} = 45.2$, $\mu_{by} = 53.6$, $\sigma_{bs} = 5.22$, $\sigma_{bv} = 1.62$, $\kappa = 3.95 \times 10^{-4}$, $p_D = 0.746$, $\Gamma = 1.01$, $\sigma_y = 4.44$. These values compare well with those obtained using the pointwise method and small discrepancies are attributable to the bias arising from the finite difference gradient approximation with c_k held constant. Algorithm 1 used $N = 1000$ and $L = 1000$ particles.

The SPSA scheme is simpler to implement than the pointwise gradient method, but it requires choice of both the sequences (a_n) and (c_n) and therefore may require more manual tuning in pilot runs. Also, the computational cost grows linearly with the number of particles.

1.8 Conclusion

The problem of estimating the number of targets and their states, and calibrating the multi–target statistical model is difficult and only approximate inference techniques are feasible (Mahler (2007)). The focus of this work was the problem of calibrating the model. For this purpose an approximation of the true marginal likelihood of the observed data, i.e. $p(\mathbf{y}_{1:T})$ where T is the final time of the recorded data, was proposed. The estimate of the model parameters was then taken to be the maximizing argument, with respect to the model parameters, of this approximation of the true marginal likelihood. A gradient ascent algorithm was used to find the model parameters that maximise the likelihood. Although the approximate likelihood function was not computable exactly, it and its gradient was estimated using SMC. The approximate likelihood function was "characterised" by sequence of a non-negative functions on E_1, and not the space $\biguplus_{k\geq 0} E_1^k$, and even a simple SMC implementation can be reasonably efficient. However, compared to "standard" SMC applications in Doucet et al. (2001), the SMC implementation was expensive. In particular the computational cost grew quadratically with the number of particles and this limited both the number of particles and the size of the data records in the numerical examples. It was demonstrated in numerical examples that the approximation to the true likelihood was reasonable as the model parameters could be inferred by maximising it. While the results are encouraging, an important issue remains to be addressed. We have not (even empirically) characterised the bias introduced by the likelihood approximation because the exact likelihood cannot be computed. A characterization of this bias appears to be a challenging problem.

Bibliography

Bar-Shalom, Y. and T. E. Fortmann (1964). *Tracking and data association*. Mathematics in science and engineering. Boston: Academic Press.

Blackman, S. and R. Popoli (Eds.) (1999). *Design and analysis of modern tracking systems*. Artech House radar library. Boston: Artech House.

Cappé, O., S. J. Godsill, and E. Moulines (2007, May). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE 96*(5), 899 –924.

Clark, D., A. T. Cemgil, P. Peeling, and S. J. Godsill (2007). Multi-object tracking of sinusoidal components in audio with the gaussian mixture probability hypothesis density filter. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Clark, D. E. and J. Bell (2006, July). Convergence results for the particle PHD filter. *IEEE Transactions on Signal Processing 54*(7), 2652–2661.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential Monte Carlo methods for Bayesian computation. In *Bayesian Statistics 8*. Oxford University Press.

Doucet, A., N. de Freitas, and N. Gordon (Eds.) (2001). *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. New York: Springer Verlag.

Johansen, A. M., S. Singh, A. Doucet, and B. Vo (2006, June). Convergence of the SMC implementation of the PHD filter. *Methodology and Computing in Applied Probability 8*(2), 265–291.

Kiefer, J. and J. Wolfowitz (1952, September). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics 23*(3), 462–466.

Kingman, J. F. C. (1993). *Poisson Processes*. Oxford Studies in Probability. Oxford University Press.

Kleinman, N. L., J. C. Spall, and D. Q. Naiman (1999). Simulation–based optimisation with stochastic approximation using common random numbers. *Management Science 45*(11), 1571–1578.

Lund, J. and E. Thonnes (2004). Perfect simulation and inference for point processes given noisy observations. *Comput. Stat. 19*(2), 317–336.

Mahler, R. (2007). *Statistical Multisource-Multitarget Information Fusion*. Artech House.

Mahler, R. P. S. (2003, October). Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 1152–1178.

Mullane, J., B. Vo, M. D. Adams, and W. S. Wijesoma (2008). A phd filtering approach to robotic mapping. In *IEEE Conf. on Control, Automation, Robotics and Vision*.

Pflug, G. (1996). *Optimization of stochastic models: the Interface between simulation and optimization*. Kluwer Academic Publishers.

Poyiadjis, G., A. Doucet, and S. S. Singh (2005). Maximum likelihood parameter estimation using particle methods. In *Proceedings of the Joint Statistical Meeting*.

Poyiadjis, G., A. Doucet, and S. S. Singh (2009). Monte carlo for computing the score and observed information matrix in state-space models with applications to parameter estimation. Technical Report CUED/F-INFENG/TR.628, Signal Processing Laboratory, Department of Engineering, University of Cambridge.

Reid, D. (1979). An algorithm for tracking multiple targets. *IEEE Transactions on AutomaticControl 24*, 843854.

Siddenblath, H. (2003). Multi-target particle filtering for the probability hypothesis density. In *Proceedings of the International Conference on Information Fusion, Cairns, Australia*, pp. 800–806.

Singh, S., B.-N. Vo., A. Baddeley, and S. Zuyev (2009). Filters for spatial point processes. *SIAM Journal on Control and Optimization 48*, 2275–2295.

Spall, J. C. (1992, March). Multivariate stochastic approximation using simultaneous perturbation gradient approximation. *IEEE Transations on Automatic Control 37*(3), 332–341.

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization* (1st ed.). Wiley-Interscience.

Storlie, C. B., C. M. Lee, J. Hannig, and D. Nychka (2009). Tracking of multiple merging and splitting targets: A statistical perspective (with discussion). *Statistica Sinica 19*(1), 152.

Vo, B. and W.-K. Ma (2006, November). The Gaussian mixture probability hypothesis density filter. *IEEE Trans. Signal Processing 54*(11), 4091–4104.

Vo, B., S. Singh, and A. Doucet (2003). Random finite sets and sequential Monte Carlo methods in multi-target tracking. In *Proceedings of the International Conference on Information Fusion, Cairns, Australia*, pp. 792–799.

Vo, B., S. Singh, and A. Doucet (2005, October). Sequential Monte Carlo methods for multitarget filtering with random finite sets. *IEEE Transactions on Aerospace and Electronic Systems 41*(4), 1224–1245.

Whiteley, N., S. Singh, and S. Godsill. Auxiliary particle implementation of the probability hypothesis density filter. *IEEE Transactions on Aerospace and Electronic Systems*. To Appear.

Yoon, J. and S. Singh (2008, September). A bayesian approach to tracking in single molecule fluorescence microscopy. Technical Report CUED/F-INFENG/TR-612, University of Cambridge. Working paper.

Zajic, T. and R. P. S. Mahler (2003). Particle-systems implementation of the PHD multitarget tracking filter. In *Proceedings of SPIE*, pp. 291–299.