1

What Are Bayesian Filtering and Smoothing?

The term *optimal filtering* traditionally refers to a class of methods that can be used for estimating the state of a time-varying system that is indirectly observed through noisy measurements. The term *optimal* in this context refers to statistical optimality. *Bayesian filtering* refers to the Bayesian way of formulating optimal filtering. In this book we use these terms interchangeably and always mean Bayesian filtering.

In optimal, Bayesian, and Bayesian optimal filtering, the *state* of the system refers to the collection of dynamic variables, such as position, velocity, orientation, and angular velocity, which fully describe the system. The *noise* in the measurements means that they are uncertain; even if we knew the true system state, the measurements would not be deterministic functions of the state but would have a distribution of possible values. The time evolution of the state is modeled as a dynamic system that is perturbed by a certain *process noise*. This noise is used for modeling the uncertainties in the system dynamics. In most cases the system is not truly stochastic, but stochasticity is used to represent the model uncertainties.

Bayesian smoothing (or optimal smoothing) is often considered to be a class of methods within the field of Bayesian filtering. While Bayesian filters in their basic form only compute estimates of the current state of the system given the history of measurements, Bayesian smoothers can be used to reconstruct states that happened before the current time. Although the term *smoothing* is sometimes used in a more general sense for methods that generate a smooth (as opposed to rough) representation of data, in the context of Bayesian filtering, the term (Bayesian) smoothing has this more definite meaning.

1.1 Applications of Bayesian Filtering and Smoothing

Phenomena that can be modeled as time-varying systems of the above type are very common in engineering applications. This kind of model can be found, for example, in navigation, aerospace engineering, space engineering, remote surveillance, telecommunications, physics, audio signal processing, control engineering, finance, and many other fields. Examples of such applications are the following.

- *Global positioning system (GPS)* (Kaplan, 1996) is a widely used satellite navigation system, where the GPS receiver unit measures arrival times of signals from several GPS satellites and computes its position based on these measurements (see Figure 1.1). The GPS receiver typically uses an extended Kalman filter (EKF) or some other optimal filtering algorithm[1] for computing the current position and velocity such that the measurements and the assumed dynamics (laws of physics) are taken into account. Also, the ephemeris information, which is the satellite reference information transmitted from the satellites to the GPS receivers, is typically generated using optimal filters.

Figure 1.1 In the GPS system, the measurements are time delays of satellite signals, and the optimal filter (e.g., extended Kalman filter, EKF) computes the position and the accurate time.

- *Target tracking* (Bar-Shalom et al., 2001; Crassidis and Junkins, 2004; Challa et al., 2011) refers to the methodology where a set of sensors,

[1] Strictly speaking, the EKF is only an approximate optimal filtering algorithm because it uses a Taylor series-based Gaussian approximation to the non-Gaussian optimal filtering solution.

such as active or passive radars, radio frequency sensors, acoustic arrays, infrared sensors, or other types of sensors, are used for determining the position and velocity of a remote target (see Figure 1.2). When this tracking is done continuously in time, the dynamics of the target and measurements from the different sensors are most naturally combined using an optimal filter or smoother. The target in this (single) target tracking case can be, for example, a robot, a satellite, a car, or an airplane.

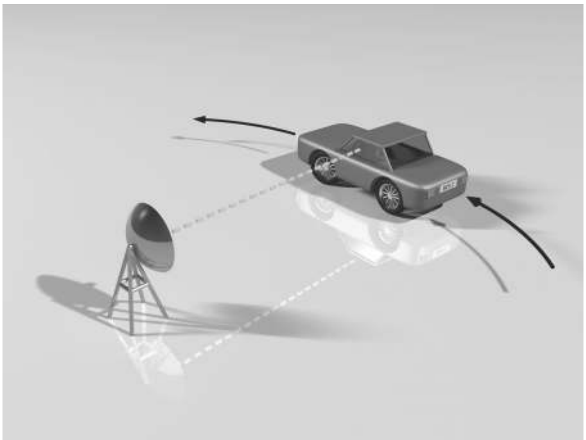

Figure 1.2 In target tracking, a sensor (e.g., radar) generates measurements (e.g., angle and distance measurements) of the target, and the purpose is to determine the target trajectory.

- *Multiple target tracking* (Bar-Shalom and Li, 1995; Blackman and Popoli, 1999; Mahler, 2014; Stone et al., 2014) systems are used for remote surveillance in cases where there are multiple targets moving at the same time in the same geographical area (see Figure 1.3). This introduces the concept of data association (which measurement was from which target?) and the problem of estimating the number of targets. Multiple target tracking systems are typically used in remote surveillance for military purposes, but their civil applications are, for example, monitoring of car tunnels, automatic alarm systems, and people tracking in buildings.
- *Inertial navigation* (Titterton and Weston, 1997; Grewal et al., 2001) uses inertial sensors, such as accelerometers and gyroscopes, for computing the position and velocity of a device, such as a car, an airplane,

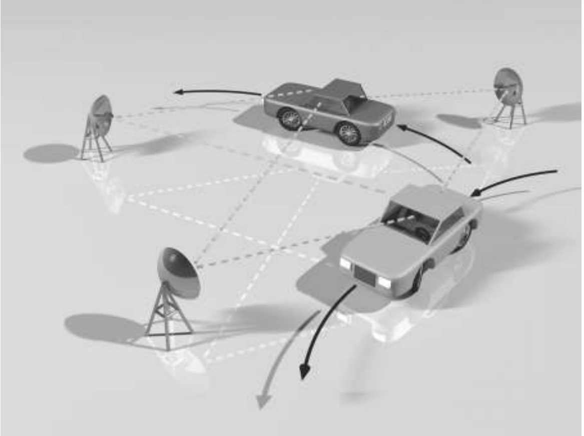

Figure 1.3 In multiple target tracking, the data association problem has to be solved because it is impossible to know without any additional information which target produced which measurement.

or a missile. When the inaccuracies in sensor measurements are taken into account, the natural way of computing the estimates is by using an optimal filter or smoother. Also, in sensor calibration, which is typically done in a time-varying environment, optimal filters and smoothers can be applied.

- *Integrated inertial navigation* (Bar-Shalom et al., 2001; Grewal et al., 2001) combines the good sides of unbiased but inaccurate sensors, such as altimeters and landmark trackers, and biased but locally accurate inertial sensors. Combination of these different sources of information is most naturally performed using an optimal filter, such as the extended Kalman filter. This kind of approach was used, for example, in the guidance system of the Apollo 11 lunar module (Eagle), which landed on the moon in 1969.
- *GPS/INS navigation* (Bar-Shalom et al., 2001; Grewal et al., 2001) is a form of integrated inertial navigation where the inertial navigation system (INS) is combined with a GPS receiver unit. In a GPS/INS navigation system, the short-term fluctuations of the GPS can be compensated by the inertial sensors, and the inertial sensor biases can be compensated by the GPS receiver. An additional advantage of this approach is that it is possible to temporarily switch to pure inertial navigation when the GPS

receiver is unable to compute its position (i.e., has no fix) for some reason. This happens, for example, indoors, in tunnels, and in other cases when there is no direct line-of-sight between the GPS receiver and the satellites.

- *Robotics and autonomous systems* (Thrun et al., 2005; Barfoot, 2017) typically use combinations of tracking and inertial navigation methods, along with sensors that measure the characteristics of the environment in one way or another. Examples of characteristics of the environment are radio signals or the locations of obstacles or landmarks detected from camera images. As the environment of the robot or autonomous system is typically unknown, the map of the environment also needs to be generated during the localization process. This concept is called simultaneous localization and mapping (SLAM), and the methodology for this purpose includes, for example, extended Kalman filters and particle filters.
- *Brain imaging* methods, such as electroencephalography (EEG), magnetoencephalography (MEG), parallel functional magnetic resonance imaging (fMRI), and diffuse optical tomography (DOT) (see Figure 1.4), are based on reconstruction of the source field in the brain from noisy sensor data by using the minimum norm estimates (MNE) technique and its variants (Hauk, 2004; Tarantola, 2004; Kaipio and Somersalo, 2005; Lin et al., 2006). The minimum norm solution can also be interpreted in the Bayesian sense as a problem of estimating the field with certain prior structure from Gaussian observations. With that interpretation, the estimation problem becomes equivalent to a *statistical inversion* or generalized *Gaussian process regression problem* (Tarantola, 2004; Kaipio and Somersalo, 2005; Rasmussen and Williams, 2006; Särkkä, 2011). Including dynamical priors then leads to a linear or non-linear spatio-temporal estimation problem, which can be solved with Kalman filters and smoothers (cf. Hiltunen et al., 2011; Särkkä et al., 2012b). The same can be done in inversion-based approaches to parallel fMRI, such as inverse imaging (InI) (Lin et al., 2006).
- *Spread of infectious diseases* (Keeling and Rohani, 2007) can often be modeled as differential equations for the number of susceptible, infected, recovered, and dead individuals. When uncertainties are introduced into the dynamic equations, and when the measurements are not perfect, the estimation of the spread of the disease can be formulated as an optimal filtering problem (see, e.g., Särkkä and Sottinen, 2008).
- *Biological processes* (Murray, 1993), such as population growth, predator–prey models, and several other dynamic processes in biology,

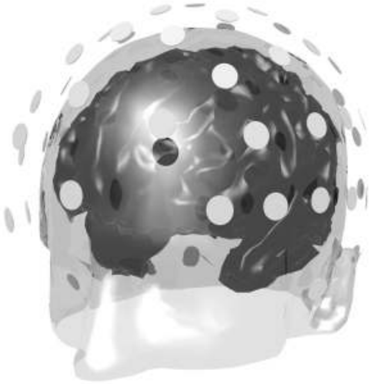

Figure 1.4 Brain imaging methods such as EEG and MEG are based on estimating the state of the brain from sensor readings. In the dynamic case, the related inversion problem can be solved with an optimal filter or smoother.

can also be modeled as (stochastic) differential equations. Estimation of the states of these processes from inaccurate measurements can be formulated as an optimal filtering and smoothing problem.

- *Telecommunications* is also a field where optimal filters are traditionally used. For example, optimal receivers, signal detectors, and phase locked loops can be interpreted to contain optimal filters (Van Trees, 1968, 1971; Proakis, 2001) as components. Also, the celebrated Viterbi algorithm (Viterbi, 1967) can be seen as a method for computing the maximum a posteriori (MAP) Bayesian smoothing solution for the underlying hidden Markov model (HMM).
- *Audio signal processing* applications, such as audio restoration (Godsill and Rayner, 1998) and audio signal enhancement (Fong et al., 2002), often use time-varying autoregressive (TVAR) models as the underlying audio signal models. These kinds of models can be efficiently estimated using optimal filters and smoothers.
- *Stochastic optimal control* (Aoki, 1967; Maybeck, 1982a; Stengel, 1994) considers control of time-varying stochastic systems. Stochastic controllers can typically be found in, for example, airplanes, cars, and rockets. Optimal, in addition to statistical optimality, means that the control

signal is constructed to minimize a performance cost, such as the expected time to reach a predefined state, the amount of fuel consumed, or the average distance from a desired position trajectory. When the state of the system is observed through a set of sensors, as it usually is, optimal filters are needed for reconstructing the state from them.

- *Learning systems* or adaptive systems can often be mathematically formulated in terms of optimal filters and smoothers (Haykin, 2001), and they have a close relationship to Bayesian non-parametric modeling, machine learning, and neural network modeling (Bishop, 2006). Methods similar to the data association methods in multiple target tracking are also applicable to on-line adaptive classification (Andrieu et al., 2002). The connection between Gaussian process regression (Rasmussen and Williams, 2006) and optimal filtering has also been discussed, for example, in Särkkä et al. (2007a), Hartikainen and Särkkä (2010), Särkkä et al. (2013), and Särkkä and Solin (2019).
- *Physical systems* that are time-varying and measured through non-ideal sensors can sometimes be formulated as stochastic state space models, and the time evolution of the system can be estimated using optimal filters (Kaipio and Somersalo, 2005). These kinds of problem are often called *inverse problems* (Tarantola, 2004), and optimal filters and smoothers can be seen as the Bayesian solutions to time-varying inverse problems.

1.2 Origins of Bayesian Filtering and Smoothing

The roots of Bayesian analysis of time-dependent behavior are in the field of optimal linear filtering. The idea of constructing mathematically optimal recursive estimators was first presented for linear systems due to their mathematical simplicity, and the most natural optimality criterion from both the mathematical and modeling points of view was least squares optimality. For linear systems, the optimal Bayesian solution (with minimum mean squared error, MMSE, loss) coincides with the least squares solution, that is, the optimal least squares solution is exactly the posterior mean.

The history of optimal filtering starts from the *Wiener filter* (Wiener, 1950), which is a frequency-domain solution to the problem of least squares optimal filtering of stationary Gaussian signals. The Wiener filter is still important in communication applications (Proakis, 2001), digital signal processing (Hayes, 1996), and image processing (Gonzalez and Woods, 2008). The disadvantage of the Wiener filter is that it can only be applied to stationary signals.

The success of optimal linear filtering in engineering applications is mostly due to the seminal article of Kalman (1960b), which describes the recursive solution to the optimal discrete-time (sampled) linear filtering problem. One reason for its success is that the *Kalman filter* can be understood and applied with very much lighter mathematical machinery than the Wiener filter. Also, despite its mathematical simplicity and generality, the Kalman filter (or actually the Kalman–Bucy filter (Kalman and Bucy, 1961)) contains the Wiener filter as its limiting special case.

In the early stages of its history, the Kalman filter was soon discovered to belong to the class of Bayesian filters (Ho and Lee, 1964; Lee, 1964; Jazwinski, 1966, 1970). The corresponding Bayesian smoothers (Rauch, 1963; Rauch et al., 1965; Leondes et al., 1970) were also developed soon after the invention of the Kalman filter. An interesting historical detail is that while Kalman and Bucy were formulating the linear theory in the United States, Stratonovich was doing the pioneering work on the probabilistic (Bayesian) approach in Russia (Stratonovich, 1968; Jazwinski, 1970).

As discussed in the book of West and Harrison (1997), in the 1960s, Kalman filter-like recursive estimators were also used in the Bayesian community, and it is not clear whether the theory of Kalman filtering or the theory of *dynamic linear models* (DLM) came first. Although these theories were originally derived from slightly different starting points, they are equivalent. Because of the Kalman filter's useful connection to the theory and history of stochastic optimal control, this book approaches the Bayesian filtering problem from the Kalman filtering point of view.

Although the original derivation of the *Kalman filter* was based on the least squares approach, the same equations can be derived from pure probabilistic Bayesian analysis. The Bayesian analysis of Kalman filtering is well covered in the classical book of Jazwinski (1970) and a bit more recently in the book of Bar-Shalom et al. (2001). Kalman filtering, mostly because of its least squares interpretation, has been widely used in stochastic optimal control. A practical reason for this is that the inventor of the Kalman filter, Rudolph E. Kalman, has also made several contributions to the theory of *linear quadratic Gaussian* (LQG) regulators (Kalman, 1960a), which are fundamental tools of stochastic optimal control (Stengel, 1994; Maybeck, 1982a).

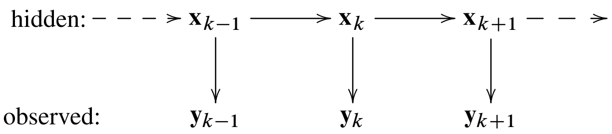

Figure 1.5 In optimal filtering and smoothing problems a sequence of hidden states $\mathbf{x}_k$ is indirectly observed through noisy measurements $\mathbf{y}_k$.

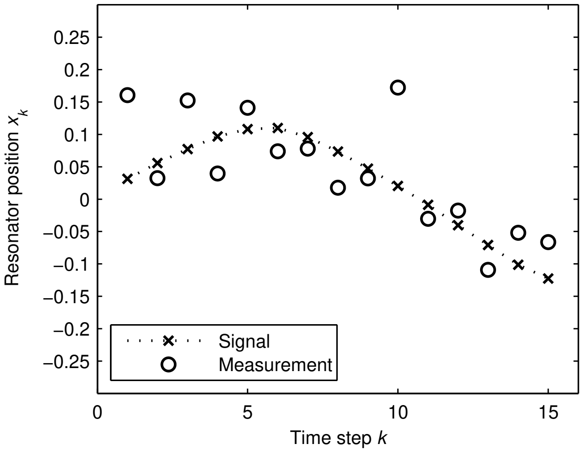

Figure 1.6 An example of a time series, which models a discrete-time resonator. The actual resonator state (signal) is hidden and only observed through the noisy measurements.

1.3 Optimal Filtering and Smoothing as Bayesian Inference

In mathematical terms, optimal filtering and smoothing are considered to be statistical inversion problems, where the unknown quantity is a vector-valued time series $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots\}$ that is observed through a set of noisy measurements $\{\mathbf{y}_1, \mathbf{y}_2, \ldots\}$, as illustrated in Figure 1.5. An example of this kind of time series is shown in Figure 1.6. The process shown is a noisy resonator with a known angular velocity. The state $\mathbf{x}_k = (x_k \; \dot{x}_k)^\mathsf{T}$ is two dimensional (2D) and consists of the position of the resonator x_k and its time derivative $\dot{x}_k$. The measurements y_k are scalar observations of the resonator position (signal), and they are corrupted by measurement noise.

The purpose of the *statistical inversion* at hand is to estimate the hidden states $\mathbf{x}_{0:T} = \{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$ from the observed measurements $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$, which means that in the Bayesian sense we want to compute the joint *posterior distribution* of all the states given all the measurements. In principle, this can be done by a straightforward application of Bayes' rule

$$p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}) = \frac{p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})\, p(\mathbf{x}_{0:T})}{p(\mathbf{y}_{1:T})}, \tag{1.1}$$

where

- $p(\mathbf{x}_{0:T})$ is the *prior distribution* defined by the dynamic model,
- $p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})$ is the likelihood model for the measurements,
- $p(\mathbf{y}_{1:T})$ is the normalization constant defined as

$$p(\mathbf{y}_{1:T}) = \int p(\mathbf{y}_{1:T} \mid \mathbf{x}_{0:T})\, p(\mathbf{x}_{0:T})\, \mathrm{d}\mathbf{x}_{0:T}. \tag{1.2}$$

Unfortunately, this full posterior formulation has the serious disadvantage that each time we obtain a new measurement, the full posterior distribution would have to be recomputed. This is particularly a problem in dynamic estimation (which is exactly the problem we are solving here!), where measurements are typically obtained one at a time, and we would want to compute the best possible estimate after each measurement. When the number of time steps increases, the dimensionality of the full posterior distribution also increases, which means that the computational complexity of a single time step increases. Thus eventually the computations will become intractable, no matter how much computational power is available. Without additional information or restrictive approximations, there is no way of getting over this problem in the full posterior computation.

However, the above problem only arises when we want to compute the *full* posterior distribution of the states at each time step. If we are willing to relax this a bit and be satisfied with selected marginal distributions of the states, the computations become an order of magnitude lighter. To achieve this, we also need to restrict the class of dynamic models to probabilistic Markov sequences, which is not as restrictive as it may at first seem. The model for the states and measurements will be assumed to be of the following type.

- **An initial distribution** specifies the *prior probability distribution* $p(\mathbf{x}_0)$ of the hidden state $\mathbf{x}_0$ at the initial time step $k = 0$.