Sigma Point and Particle Approximations of Stochastic Differential Equations in Optimal Filtering

Simo Särkkä

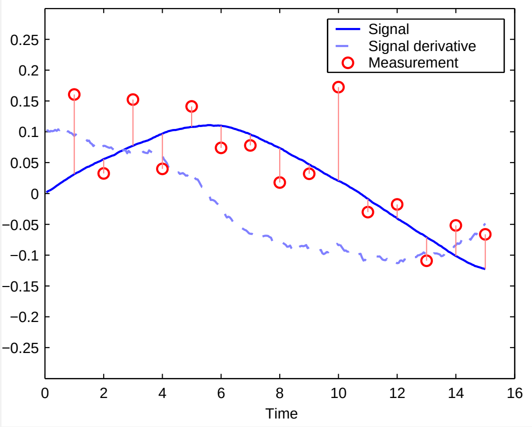
Helsinki University of Technology
Nalco Company

May, 2008

Contents

Continuous-Discrete Filtering Problem

- Estimate the unobserved continuous-time signal from noisy discrete-time measurements



Mathematical Problem Formulation

- The dynamics of state $\mathbf{x}(t)$ modeled as a stochastic differential equation (Itô diffusion)

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,dt + \mathbf{L}\,d\beta(t).$$

- Measurements $\mathbf{y}_k$ are obtained at discrete times

$$\mathbf{y}_k \sim p(\mathbf{y}_k \,|\, \mathbf{x}(t_k)).$$

- Formal solution: Compute the posterior distribution(s)

$$p(\mathbf{x}(t) \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_k), \qquad t \geq t_k.$$

Formal solution

Optimal filter

1. Prediction step: Solve the Kolmogorov-forward (Fokker-Planck) partial differential equation.

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i} (f_i(\mathbf{x}, t)\, p) + \frac{1}{2} \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \left([\mathbf{L}\, \mathbf{Q}\, \mathbf{L}^T]_{ij}\, p \right)$$

2. Update step: Apply the Bayes' rule.

$$p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k}) = \frac{p(\mathbf{y}_k \,|\, \mathbf{x}(t_k))\, p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k-1})}{\int p(\mathbf{y}_k \,|\, \mathbf{x}(t_k))\, p(\mathbf{x}(t_k) \,|\, \mathbf{y}_{1:k-1})\, d\mathbf{x}(t_k)}$$

Probability Density Approximations

- Common types of probability density approximations:
  - Gaussian approximations: Taylor series, statistical linearization, unscented transform (UT).
  - Parametric PDF models: assumed density, mixture models, variational approximations.
  - Monte Carlo approximations: perfect Monte Carlo sampling, importance sampling, Markov chain Monte Carlo (MCMC).
- Here we shall concentrate on the following methods:
  - Continuous-discrete extended Kalman filter (EKF), which is a Taylor/Gaussian approximation based method.
  - Continuous-discrete unscented Kalman filter (UKF), which is a UT/Gaussian approximation based method.
  - Continuous-discrete sequential importance resampling (SIR), which is an importance sampling based Monte Carlo method.

Taylor Series Approximations of Transformations

- Consider transformation of Gaussian random variable by non-linear function $\mathbf{g}(\cdot)$:

$$\mathbf{x} \sim \mathrm{N}(\mathbf{m}, \mathbf{P})$$
$$\mathbf{y} = \mathbf{g}(\mathbf{x})$$

- The function can be approximated by Taylor series:

$$\mathbf{g}(\mathbf{m} + \Delta\mathbf{x}) = \mathbf{g}(\mathbf{m}) + \mathbf{G}(\mathbf{m})\,\Delta\mathbf{x} + \ldots$$

  where $\mathbf{G}(\cdot)$ is the Jacobian of $\mathbf{g}(\cdot)$.
- We get the following Gaussian approximation to the distribution of the random variable $\mathbf{y}$:

$$\mathbf{y} \sim \mathrm{N}\left(\mathbf{g}(\mathbf{m}), \mathbf{G}(\mathbf{m})\,\mathbf{P}\,\mathbf{G}^T(\mathbf{m})\right)$$

Extended Kalman Filter (EKF)

- EKF applies the Taylor series approximation to the filtering model

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{q}, \qquad \mathbf{q} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q})$$
$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k) + \mathbf{r}, \qquad \mathbf{r} \sim \mathrm{N}(\mathbf{0}, \mathbf{R})$$

- The resulting EKF equations are of the form:
  - *Prediction:*

$$\mathbf{m}_k^- = \mathbf{f}(\mathbf{m}_{k-1})$$
$$\mathbf{P}_k^- = \mathbf{F}(\mathbf{m}_{k-1})\, \mathbf{P}_{k-1}\, \mathbf{F}^T(\mathbf{m}_{k-1}) + \mathbf{Q}.$$

  - *Update:*

$$\mathbf{S}_k = \mathbf{H}(\mathbf{m}_k^-)\, \mathbf{P}_k^-\, \mathbf{H}^T(\mathbf{m}_k^-) + \mathbf{R}$$
$$\mathbf{K}_k = \mathbf{P}_k^-\, \mathbf{H}^T(\mathbf{m}_k^-)\, \mathbf{S}_k^{-1}$$
$$\mathbf{m}_k = \mathbf{m}_k^- + \mathbf{K}_k\, [\mathbf{y}_k - \mathbf{h}(\mathbf{m}_k^-)]$$
$$\mathbf{P}_k = \mathbf{P}_k^- - \mathbf{K}_k\, \mathbf{S}_k\, \mathbf{K}_k^T.$$

Continuous-Discrete EKF [1/2]

- In continuous-discrete filtering the dynamic model is a stochastic differential equation (SDE):

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})\, dt + \mathbf{L}\, d\beta$$

- Taking first order discrete-time approximation we get (note that the SDE can be interpreted as a Stratonovich equation)

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{f}(\mathbf{x}_{k-1})\, \delta t + \mathbf{q}, \qquad \mathbf{q} \sim \mathrm{N}(\mathbf{0}, \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^T\, \delta t)$$

- Continuous solution on interval $[0, T]$ (between measurements) can be approximated by iterating the discrete approximation over the interval in steps of δt.

Continuous-Discrete EKF [2/2]

- The EKF prediction equations up to first order in δt:

$$
\begin{aligned}
\mathbf{m}_k &= \mathbf{m}_{k-1} + \mathbf{f}(\mathbf{m}_{k-1})\,\delta t \\
\mathbf{P}_k &= \mathbf{P}_{k-1} + \mathbf{F}(\mathbf{m}_{k-1})\,\mathbf{P}_{k-1}\,\delta t + \mathbf{P}_{k-1}\,\mathbf{F}^T(\mathbf{m}_{k-1})\,\delta t + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^T\,\delta t
\end{aligned}
$$

- Dividing by δt and by taking limit $\delta t \to 0$, we get

$$
\begin{aligned}
\frac{d\mathbf{m}}{dt} &= \mathbf{f}(\mathbf{m}) \\
\frac{d\mathbf{P}}{dt} &= \mathbf{F}(\mathbf{m})\,\mathbf{P} + \mathbf{P}\,\mathbf{F}^T(\mathbf{m}) + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^T
\end{aligned}
$$

- These differential equations are satisfied between measurements.
- At measurements we use discrete-time EKF update equations.

Unscented Transform (UT) [1/2]

- The unscented transform also considers transformations as

$$\mathbf{x} \sim \mathrm{N}(\mathbf{m}, \mathbf{P})$$
$$\mathbf{y} = \mathbf{g}(\mathbf{x})$$

- Instead of the Taylor series, a set of sigma points are computed as the columns of the Cholesky factorization of **P**:

$$\begin{aligned}
\mathbf{x}^{(0)} &= \mathbf{m} \\
\mathbf{x}^{(i)} &= \mathbf{m} + c \left[\sqrt{\mathbf{P}}\right]_i, \quad i = 1, \ldots, n \\
\mathbf{x}^{(i)} &= \mathbf{m} - c \left[\sqrt{\mathbf{P}}\right]_i, \quad i = n+1, \ldots, 2n
\end{aligned}$$

Unscented Transform (UT) [2/2]

- The sigma points are then propagated through the function $\mathbf{g}(\cdot)$:

$$\mathbf{y}^{(i)} = \mathbf{g}(\mathbf{x}^{(i)}), \quad i = 0, \ldots, 2n.$$

- The mean and covariance of $\mathbf{y}$ are approximated as linear combinations of the resulting points:

$$\begin{aligned} \boldsymbol{\mu} &\approx \sum_{i=0}^{2n} W_i^{(m)} \, \mathbf{y}^{(i)} \\ \mathbf{S} &\approx \sum_{i=0}^{2n} W_i^{(c)} \, (\mathbf{y}^{(i)} - \boldsymbol{\mu}) \, (\mathbf{y}^{(i)} - \boldsymbol{\mu})^T. \end{aligned}$$

- The sigma points are chosen deterministically and the weights are fixed and thus this is not a Monte Carlo approach.

Unscented Kalman Filter

- The unscented Kalman filter (UKF) is almost like an EKF, but uses unscented transforms instead of Taylor series expansions.
- The prediction and update equations are messy, but the idea is the following:
  - *Prediction step:*
    1. Form sigma points of the state $\mathbf{x}_{k-1}$
    2. Propagate them through the dynamic model function
    3. Compute the resulting mean and covariance
    4. Add the process noise covariance to state covariance
  - *Update step:*
    1. Form sigma points of the predicted state
    2. Form UT approximation of the joint distribution of predicted state and measurement
    3. Use computation rules of Gaussian distributions for conditioning the joint distribution to the measurement $\mathbf{y_k}$

Matrix Form of Unscented Transform

- Define matrices of sigma points as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(0)} & \cdots & \mathbf{x}^{(2n)} \end{bmatrix}$$
$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}^{(0)} & \cdots & \mathbf{y}^{(2n)} \end{bmatrix}$$

- Propagation of sigma points can be written as matrix operation

$$\mathbf{Y} = \mathbf{g}(\mathbf{X})$$

- The mean and covariance computation equations can be written as matrix expressions

$$\mu \approx \mathbf{Y}\,\mathbf{w}_m$$
$$\mathbf{S} \approx \mathbf{Y}\,\mathbf{W}\,\mathbf{Y}^T$$

where $\mathbf{w}_m$ and $\mathbf{W}$ are constant vector and matrix.

Matrix Form of Unscented Kalman Filter

- Unscented Kalman filter can be written in matrix form:
  - *Prediction:*

$$
\begin{aligned}
\mathbf{X}_{k-1} &= \begin{bmatrix}\mathbf{m}_{k-1} & \cdots & \mathbf{m}_{k-1}\end{bmatrix} + c\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}_{k-1}} & -\sqrt{\mathbf{P}_{k-1}}\end{bmatrix} \\
\mathbf{m}_k^- &= \mathbf{f}(\mathbf{X}_{k-1})\,\mathbf{w}_m \\
\mathbf{P}_k^- &= \mathbf{f}(\mathbf{X}_{k-1})\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}_{k-1}) + \mathbf{Q}.
\end{aligned}
$$

  - *Update:*

$$
\begin{aligned}
\mathbf{X}_k^- &= \begin{bmatrix}\mathbf{m}_k^- & \cdots & \mathbf{m}_k^-\end{bmatrix} + c\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}_k^-} & -\sqrt{\mathbf{P}_k^-}\end{bmatrix} \\
\mathbf{S}_k &= \mathbf{h}(\mathbf{X}_k^-)\,\mathbf{W}\,\mathbf{h}^T(\mathbf{X}_k^-) + \mathbf{R} \\
\mathbf{K}_k &= \mathbf{X}_k^-\,\mathbf{W}\,\mathbf{h}^T(\mathbf{X}_k^-)\,\mathbf{S}_k^{-1} \\
\mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\left[\mathbf{y}_k - \mathbf{h}(\mathbf{X}_k^-)\,\mathbf{w}_m\right] \\
\mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^T.
\end{aligned}
$$

Continuous-Discrete UKF [1/2]

- Taking the continuous-time limit of the prediction step leads to the equations:

$$
\begin{aligned}
\mathbf{X} &= \begin{bmatrix}\mathbf{m} & \cdots & \mathbf{m}\end{bmatrix} + c \begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{P}} & -\sqrt{\mathbf{P}}\end{bmatrix} \\
\frac{d\mathbf{m}}{dt} &= \mathbf{f}(\mathbf{X})\,\mathbf{w}_m \\
\frac{d\mathbf{P}}{dt} &= \mathbf{X}\,\mathbf{W}\,\mathbf{f}^T(\mathbf{X}) + \mathbf{f}(\mathbf{X})\,\mathbf{W}\,\mathbf{X}^T + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^T
\end{aligned}
$$

- In continuous-discrete UKF the above differential equations are used between the measurements.
- The discrete-time UKF update equations are used at measurements times.
- The matrix UT computations can be replaced with corresponding summation formulas, which are computationally lighter.

Continuous-Discrete UKF [2/2]

- The continuous UKF prediction equations can be written in terms of sigma points as

$$\mathbf{M} = \mathbf{A}^{-1} \, [\mathbf{X} \, \mathbf{W} \, \mathbf{f}^T(\mathbf{X}) + \mathbf{f}(\mathbf{X}) \, \mathbf{W} \, \mathbf{X}^T + \mathbf{L} \, \mathbf{Q} \, \mathbf{L}^T] \, \mathbf{A}^{-T}$$

$$\frac{d\mathbf{X}_i}{dt} = \mathbf{f}(\mathbf{X}, t) \, \mathbf{w}_m + c \, \begin{bmatrix} \mathbf{0} & \mathbf{A} \, \Phi\left(\mathbf{M}\right) & -\mathbf{A} \, \Phi\left(\mathbf{M}\right) \end{bmatrix}_i$$

- The matrix **A** is the Cholesky factor of **P**, which can be found by collecting suitable terms from **X** and by subtracting the mean.
- $\Phi(\cdot)$ is a function returning the lower diagonal part of the argument as follows:

$$\Phi_{ij}\left(\mathbf{M}\right) = \begin{cases} M_{ij} & , \quad \text{if } i > j \\ \frac{1}{2} M_{ij} & , \quad \text{if } i = j \\ 0 & , \quad \text{if } i < j. \end{cases}$$

Sequential Importance Resampling

Sequential Importance Resampling

1. Draw a random sample from the importance distribution

$$\mathbf{x}^{(i)}(t_k) \sim q(\mathbf{x}^{(i)}(t_k) \,|\, \mathbf{x}^{(i)}(t_{k-1}))$$

2. Evaluate the importance weight

$$w_k^{(i)} \propto \frac{p(\mathbf{y}_k \,|\, \mathbf{x}^{(i)}(t_k))\, p(\mathbf{x}^{(i)}(t_k) \,|\, \mathbf{x}^{(i)}(t_{k-1}))}{q(\mathbf{x}^{(i)}(t_k) \,|\, \mathbf{x}^{(i)}(t_{k-1}))}$$

3. Do resampling if needed.

The Problem of SIR Weight Evaluation

- The weight evaluation of SIR is of the form

$$w_k^{(i)} \propto \frac{p(\mathbf{y}_k \,|\, \mathbf{x}^{(i)}(t_k))\, p(\mathbf{x}^{(i)}(t_k) \,|\, \mathbf{x}^{(i)}(t_{k-1}))}{q(\mathbf{x}^{(i)}(t_k) \,|\, \mathbf{x}^{(i)}(t_{k-1}))}$$

- But $p(\mathbf{x}(t_k) \,|\, \mathbf{x}(t_{k-1}))$ is the solution of an arbitrary second order partial differential equation and cannot be solved.
- Actually we only need the likelihood ratio

$$\frac{p(\mathbf{x}(t_k) \,|\, \mathbf{x}(t_{k-1}))}{q(\mathbf{x}(t_k) \,|\, \mathbf{x}(t_{k-1}))}$$

- This can be computed with the Girsanov theorem without solving the PDE.

Girsanov Theorem

- Let $\theta(t)$ be a stochastic process, which is driven by ("adapted to") a Brownian motion $\beta(t)$.
- The likelihood ratio between $\theta(t)$ and $\beta(t)$ is:

$$\frac{dP_{\theta}}{dP_{\beta}} = \exp\left(\int_0^t \theta^T(t)\, d\beta(t) - \frac{1}{2} \int_0^t ||\theta(t)||^2\, dt \right).$$

- The likelihood ratio can be exactly computed by above stochastic integral.
- Efficient simulation based numerical solutions possible.

Evaluating the Likelihood Ratio

- With Girsanov theorem, we can derive expression for likelihood ratio for two SDE's:

$$\begin{aligned} d\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\, dt + \mathbf{L}\, d\beta \\ d\mathbf{s} &= \mathbf{g}(\mathbf{s}, t)\, dt + \mathbf{B}\, d\beta. \end{aligned}$$

- Process $\mathbf{s}(t)$ can be the importance process for estimated process $\mathbf{x}(t)$.
- It is a *stochastic integral*: Well known numerical methods for SDE's can be used.
- It is a *Monte Carlo solution*: Solution converges to the exact solution.

Non-invertibility of Diffusion Matrix [1/3]

- In mathematical analysis of SDEs it is often assumed that in the model

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\, dt + \mathbf{L}\, d\beta$$

  matrix $\mathbf{L}$ is square, invertible or even scalar.
- This assumption eases the mathematical analysis, and is a feasible assumption, for example, in models of stock prices.
- But in physics based models, $\mathbf{L}$ is almost never invertible.
- In this case the noise term $\mathbf{L}\, d\beta$ is singular in the sense that its diffusion matrix $\mathbf{L}\, \mathbf{Q}\, \mathbf{L}^T$ is singular.

Non-invertibility of Diffusion Matrix [2/3]

- For example, the Newton's law with white noise force:

$$\frac{d^2 x}{dt^2} = w(t)$$

- The model is equivalent to SDE

$$d \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} dt + \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{\mathbf{L}} d\beta$$

- In this model $\mathbf{L}$ is not square and $\mathbf{L}\,\mathbf{L}^T$ is not invertible.
- The same problem always arises, when some state component is continuously differentiable with respect to time (almost always in physics).

Non-invertibility of Diffusion Matrix [3/3]

- The non-invertibility of the diffusion matrix is not an issue to continuous-discrete EKF or UKF.
- But Girsanov theorem has a problem with this, because the process is no longer absolutely continuous with respect to any Brownian motion.
- Fortunately, by directly computing the likelihood ratio between two processes with similar singularities, this problem can be avoided.
- This way likelihood ratio based particle filter approach can be generalized to the singular types of diffusion models also.

Rao-Blackwellization [1/2]

- Sometimes, the dynamic model is conditionally linear Gaussian as follows:

$$
\begin{aligned}
d\mathbf{x} &= \mathbf{F}(\mathbf{s})\,\mathbf{x}\,dt + \mathbf{L}\,d\beta \\
d\mathbf{s} &= \mathbf{g}(\mathbf{s})\,dt + \mathbf{B}\,d\eta
\end{aligned}
$$

- Given the process **s** the process **x** is a Gaussian process.
- The Brownian motion in the first equation can be now marginalized out (Rao-Blackwellized), which leads to the model

$$
\begin{aligned}
d\mathbf{m}/dt &= \mathbf{F}(\mathbf{s})\,\mathbf{m} \\
d\mathbf{P}/dt &= \mathbf{F}(\mathbf{s})\,\mathbf{P} + \mathbf{P}\,\mathbf{F}^T(\mathbf{s}) + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^T \\
d\mathbf{s} &= \mathbf{g}(\mathbf{s})\,dt + \mathbf{B}\,d\eta
\end{aligned}
$$

Rao-Blackwellization [2/2]

- If the measurement model is also suitably conditionally linear Gaussian, we may apply the Kalman filter update equations on the measurement step.
- This leads to a Rao-Blackwellized particle filtering algorithm, where part of the state components are replaced with their sufficient statistics.
- Static parameters in dynamic or measurements models can be sometimes handled in a similar manner.

Toy Example: Noisy Simple Pendulum Problem

- Model of noisy simple pendulum:

$$\frac{d^2 x}{dt^2} + a^2 \sin(x) = w(t).$$

- In Brownian motion notation:

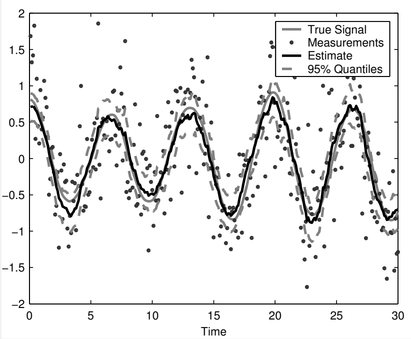
$$\begin{aligned} \frac{dx_1}{dt} &= x_2 \\ dx_2 &= -a^2 \sin(x_1)\, dt + d\beta, \end{aligned}$$

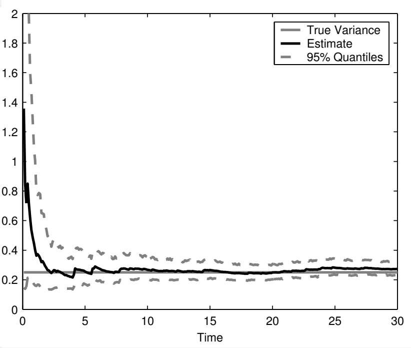
- Measurements:

$$\begin{aligned} y_k &\sim \mathrm{N}(x_1(t_k), \sigma^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \end{aligned}$$

Toy Example: Simulation Results

Evolution of signal estimate (left) and variance estimate (right):





Applications of Methods [1/2]

- Multiple target tracking (remote surveillance)
  - Target dynamics are modeled with stochastic differential equations.
  - The measurements arrive at irregular intervals.
  - The number of targets in unknown.
  - Data association indicators are also unknown latent variables.
  - Rao-Blackwellization can be typically applied.
- Bus and bus stop tracking
  - Bus dynamics are modeled with stochastic differential equations.
  - Measurements from GPS, odometer, gyroscope and acceleration sensors.
  - The index of the current bus stop is an unknown variable to be estimated.
  - The known order of bus stops is used as additional information.
  - Rao-Blackwellization can be typically applied.

Applications of Methods [2/2]

- Online paper formation estimation
  - Paper web dynamics are modeled with stochastic differential equations.
  - The sensor is moving and thus only a small part of the sheet is measured at a time.
  - Rao-Blackwellization can be typically applied, and often the process is even linear.
- Monitoring of chemical processes
  - Reaction kinetics are modeled with stochastic differential equations.
  - The measurements can be highly non-linear functions of the state.
  - Processes typically contain unknown physical parameters.

Summary

- *Continuous-discrete EKF*:
  - Taylor series based Gaussian approximation to the SDE.
  - Mean and covariance differential equations on prediction step.
- *Continuous-discrete UKF*:
  - Unscented transform instead of the Taylor series.
  - Mean and covariance differential equations on prediction step.
  - Alternatively, differential equation for the sigma-points.
- *The Girsanov theorem*:
  - Can be used for evaluating likelihood ratios of SDEs in sequential importance sampling.
  - Non-invertible diffusion matrices need special care.
  - Conditionally linear Gaussian processes can be marginalized out, which leads to Rao-Blackwellized filters.
- The methods have applications in many areas, for example, in navigation, paper industry and chemical industry.