Reasoning with Uncertainty

Estimation and Filtering

Optimal Estimation

- The goal of optimal estimation is to determine the *best* estimate of the state of the system given a set of observations
  - *Best* implies minimum error
- There are 3 general types of estimation problems that differ in terms of the available observations
  - Filtering: Determine the best estimate for the current point in time
  - Smoothing: Determine the best estimate for a point in time in the past
  - Prediction: Determine the best estimate for a point in time in the future

Probabilistic Reasoning Over Time

- Stochastic processes can be represented in terms of conditional probabilities
  - State of the system at time t: $s_t \in S$
  - Observation of the system at time t: $o_t \in O$
  - System model: $P(s_t | s_{t-1}, o_t, \ldots o_1, s_0)$
  - Observation model: $P(o_t | s_t, o_{t-1}, \ldots o_1, s_0)$
- Useful properties for stochastic processes
  - Stationarity – The process itself does not change over time
  - Markov – The state of the system depends only on a finite history (first order: only on the last state)

Dynamic Bayesian Networks

- Stochastic processes that are Markov (any order) can be represented using Dynamic Bayesian Networks
  - Replicated networks for the state at different time steps
  - Connections between time copies encode transition probabilities
  - Connections from state-related notes to observation-related nodes represent the observation model

Bayesian Filtering

- A Bayesian filter computes the posterior distribution of the state using the observations
  - Discrete case:

$$P(s_t \mid o_t, o_{t-1}, \ldots, o_1) = \frac{P(o_t \mid s_t, o_{t-1}, \ldots, o_1) P(s_t \mid o_{t-1}, \ldots, o_1)}{P(o_t \mid o_{t-1}, \ldots, o_1)}$$

  - Continuous case:

$$p(s_t \mid o_t, o_{t-1}, \ldots, o_1) = \frac{p(o_t \mid s_t, o_{t-1}, \ldots, o_1) p(s_t \mid o_{t-1}, \ldots, o_1)}{p(o_t \mid o_{t-1}, \ldots, o_1)}$$

Bayesian Filtering

- A Bayesian filter computes the posterior distribution of the state using the observations
  - Discrete case:

$$P(s_t \mid o_t, o_{t-1}, \ldots, o_1) = \frac{P(o_t \mid s_t, o_{t-1}, \ldots, o_1) P(s_t \mid o_{t-1}, \ldots, o_1)}{P(o_t \mid o_{t-1}, \ldots, o_1)}$$

$$= \frac{P(o_t \mid s_t, o_{t-1}, \ldots, o_1) \sum_{s_{t-1}} P(s_t \mid s_{t-1}, o_{t-1}, \ldots, o_1) P(s_{t-1} \mid o_{t-1}, \ldots, o_1)}{\sum_{s_{t-1}} P(o_t \mid s_{t-1}, o_{t-1}, \ldots, o_1) P(s_{t-1} \mid o_{t-1}, \ldots, o_1)}$$

  - Continuous case:

$$p(s_t \mid o_t, o_{t-1}, \ldots, o_1) = \frac{p(o_t \mid s_t, o_{t-1}, \ldots, o_1) p(s_t \mid o_{t-1}, \ldots, o_1)}{p(o_t \mid o_{t-1}, \ldots, o_1)}$$

$$= \frac{p(o_t \mid s_t, o_{t-1}, \ldots, o_1) \int_{s_{t-1}} p(s_t \mid s_{t-1}, o_{t-1}, \ldots, o_1) p(s_{t-1} \mid o_{t-1}, \ldots, o_1)\, ds_{t-1}}{\int_{s_{t-1}} p(o_t \mid s_{t-1}, o_{t-1}, \ldots, o_1) p(s_{t-1} \mid o_{t-1}, \ldots, o_1) ds_{t-1}}$$

Recursive Bayesian Filtering

- If the process is Markov the recursive Bayesian filter can be derived
  - Discrete case:

$$P(s_t \mid o_t, o_{t-1}, \ldots, o_1) = \frac{P(o_t \mid s_t, o_{t-1}, \ldots, o_1) \sum_{s_{t-1}} P(s_t \mid s_{t-1}, o_{t-1}, \ldots, o_1) P(s_{t-1} \mid o_{t-1}, \ldots, o_1)}{\sum_{s_{t-1}} P(o_t \mid s_{t-1,} o_{t-1}, \ldots, o_1) P(s_{t-1} \mid o_{t-1}, \ldots, o_1)}$$

$$= \frac{P(o_t \mid s_t) \sum_{s_{t-1}} P(s_t \mid s_{t-1}) P(s_{t-1} \mid o_{t-1}, \ldots, o_1)}{\sum_{s_{t-1}} P(o_t \mid s_{t-1}) P(s_{t-1} \mid o_{t-1}, \ldots, o_1)} = \alpha\, P(o_t \mid s_t) \sum_{s_{t-1}} P(s_t \mid s_{t-1}) P(s_{t-1} \mid o_{t-1}, \ldots, o_1)$$

  - Continuous case:

$$p(s_t \mid o_t, o_{t-1}, \ldots, o_1) = \frac{p(o_t \mid s_t, o_{t-1}, \ldots, o_1) \int_{s_{t-1}} p(s_t \mid s_{t-1}, o_{t-1}, \ldots, o_1)\, p(s_{t-1} \mid o_{t-1}, \ldots, o_1)\, ds_{t-1}}{\int_{s_{t-1}} p(o_t \mid s_{t-1,} o_{t-1}, \ldots, o_1)\, p(s_{t-1} \mid o_{t-1}, \ldots, o_1) ds_{t-1}}$$

$$= \frac{p(o_t \mid s_t) \int_{s_{t-1}} p(s_t \mid s_{t-1})\, p(s_{t-1} \mid o_{t-1}, \ldots, o_1)\, ds_{t-1}}{\int_{s_{t-1}} p(o_t \mid s_{t-1})\, p(s_{t-1} \mid o_{t-1}, \ldots, o_1) ds_{t-1}} = \alpha\, p(o_t \mid s_t) \int_{s_{t-1}} p(s_t \mid s_{t-1})\, p(s_{t-1} \mid o_{t-1}, \ldots, o_1)\, ds_{t-1}$$

Recursive Bayesian Filtering

- The recursive Bayesian filter can be broken into two phases
  - Prediction:

$$p(s_t \mid o_{t-1},\ldots,o_1) = \int_{s_{t-1}} p(s_t \mid s_{t-1})\, p(s_{t-1} \mid o_{t-1},\ldots,o_1)\, ds_{t-1}$$

  - Measurement:

$$p(s_t \mid o_t, o_{t-1},\ldots,o_1) = \frac{p(o_t \mid s_t)}{p(o_t \mid o_{t-1},\ldots,o_1)}\, p(s_t \mid o_{t-1},\ldots,o_1)$$

Recursive Bayesian Filtering

- Benefits of a Bayesian filter
  - Optimal estimates
  - No assumptions about distributions
  - Uniform framework
- Problems of the filter
  - Often computationally intractable
  - Integral might not be analytically solvable

Kalman Filter

- The Kalman filter is a special case of the recursive Bayesian filter for the following assumptions:
  - The system and observation model are linear

    $$s_t = As_{t-1} + w_t$$

    $$o_t = Hs_t + v_t$$

  - The prior distribution and the uncertainty in the system and observation models are Gaussian

    $$w_t \sim N(0, Q)$$

    $$v_t \sim N(0, R)$$

Kalman Filter

- The Kalman filter estimates the posterior distribution in terms of the mean and the Covariance matrix

$$\hat{s}_t = E[s_t]$$

$$P_t = E[(s_t - \hat{s}_t)(s_t - \hat{s}_t)^T]$$

  - The posterior distribution is a Gaussian distribution (maintaining the first two moments of the distribution)

Discrete Kalman Filter

- The discrete Kalman filter is a special version of the recursive Bayesian filter
  - Prediction:

$$\hat{s}_t^- = A\hat{s}_{t-1}$$

$$P_t^- = AP_{t-1}A^T + Q$$

  - Measurement:

$$\hat{s}_t = \hat{s}_t^- + K_t(o_t - H\hat{s}_t^-)$$

$$K_t = P_t^- H^T (HP_t^- H^T + R)^{-1}$$

$$P_t = (I - K_t H)P_t^-$$

The Kalman Gain – Example Derivation

- The Kalman gain K_t is the weight term that minimizes the expected squared difference between the estimate and the true state.
  - Derivation of K_t for a simple example:
    - The state is one-dimensional: $s_t \in \Re$, $P_t = \sigma_t^2$
    - The process is stationary: $A = 1$, $Q=0$
    - The system directly observes the state: $H=1$, $R= \sigma_o^2$
    - The prior distribution is Normal with a mean of s_0 and a variance of P_0

  Since the system is linear and all distributions are Gaussian, the resulting posterior distribution after every recursive step is a Gaussian with mean $\hat{s}_t$ and variance P_t

The Kalman Gain – Example Derivation

- Prediction:
  - The process is stationary and there is no uncertainty added at every step:

$$\hat{s}_t^- = \hat{s}_{t-1}$$

$$P_t^- = P_{t-1}$$

- Measurement:
  - Since both distributions are Gaussian:

$$\hat{s}_t = E[s_t] = K_1 \hat{s}_t^- + K_2 o_t$$

$$P_t = E[(\hat{s}_t - s_t)^2]$$

The Kalman Gain – Example Derivation

- The true state s_t is related to the estimate as in:

  $\hat{s}_t = s_t + \hat{e}_t \quad , \; E[\hat{e}_t] = P_t$

  $\hat{s}_t^- = s_t + \hat{e}_t^- \; , \; E[\hat{e}_t^-] = P_t^-$

- Using this, the goal is to find the gains K_1 and K_2 that minimize the expected value of the squared posterior error, $E[\hat{e}_t^2]$.

  $$\hat{e}_t = \hat{s}_t - s_t = (K_1\hat{s}_t^- + K_2 o_t) - s_t = K_1(s_t + \hat{e}_t^-) + K_2 o_t - s_t$$

- Since the observation is directly of the state:

  $o_t = s_t + e_o$

  $\Rightarrow$

  $\hat{e}_t = K_1(s_t + \hat{e}_t^-) + K_2(s_t + e_o) - s_t = s_t(K_1 + K_2 - 1) + K_1\hat{e}_t^- + K_2 e_o$

The Kalman Gain – Example Derivation

- In order for the estimated posterior to be unbiased, the expected value of the error has to be *0*:

$$E[\hat{e}_t] = E[s_t(K_1 + K_2 - 1) + K_1\hat{e}_t^- + K_2 e_o] = s_t(K_1 + K_2 - 1) \hat{=} 0$$

$$\Rightarrow K_2 = 1 - K_1$$

- Given this, the expected value of the posterior error is:

$$E[\hat{e}_t^2] = E[(K_1\hat{e}_t^- + (1-K_1)e_o)^2] = E[K_1^2\hat{e}_t^{i\,2} + (1-K_1)^2 e_o^2 + 2K_1(1-K_1)\hat{e}_t^- e_o]$$

- Since the state and observation errors are both *b*-mean and independently distributed:

$$E[\hat{e}_t^2] = E[K_1^2\hat{e}_t^{-2}] + E[(1-K_1)^2 e_o^2] = K_1^2 E[\hat{e}_t^{-2}] + (1-K_1)^2 E[e_o^2] = K_1^2 P_t^- + (1-K_1)^2\sigma_o^2$$

- To minimize this we set the derivative to *0* :

$$\frac{\partial E[\hat{e}_t^2]}{\partial K_1} = 2K_1P_t^- + 2(1-K_1)(-1)\sigma_o^2 = K_1(2P_t^- + 2\sigma_o^2) - 2\sigma_o^2 \hat{=} 0$$

$$\Rightarrow K_1 = \frac{\sigma_o^2}{P_t^- + \sigma_o^2} \quad , \quad \hat{s}_t = \frac{\sigma_o^2}{P_t^- + \sigma_o^2}\hat{s}_t^- + \frac{P_t^-}{P_t^- + \sigma_o^2}o_t \quad , \quad P_t = E[\hat{e}_t^2] = \left(\frac{\sigma_o^2}{P_t^- + \sigma_o^2}\right)^2 P_t^- + \left(\frac{P_t^-}{P_t^- + \sigma_o^2}\right)^2 \sigma_o^2 = \frac{\sigma_o^2 P_t^-}{P_t^- + \sigma_o^2}$$

Discrete Kalman Filter

- The discrete Kalman filter provides the optimal estimate for the posterior probability distribution given the conditions are met.
  - Always converges to the optimal estimate
  - The best estimate for the next state is usually extracted as the mean of the distribution as it minimizes multiple error metrics, e.g.:
    - Maximum likelihood estimate
    - Minimum squared error estimate

The Extended Kalman Filter

- The Extended Kalman Filter (EKF) relaxes the requirement on linear models
  - Uses the Jacobian matrix as a locally linear approximation of the function.
  - Note: The EKF does not always converge to the correct solution

Kalman Filters

- Kalman filters give optimal estimates for cases where the distributions for the estimates and the observations are Gaussian
  - Advantages
    - Optimal estimates
    - Fast filter updates: $O(1)$
  - Disadvantages
    - Only normal distributions (i.e. only unimodal estimates)
    - EKF has no optimal convergence guarantees

Discretized Bayesian Filters

- Approximate filters for non-Gaussian scenarios can be created by discretizing the state space for the distribution
  - Complexity: $O(n^2)$: n = number of state partitions

Sampling-Based Filters

- General distributions can be approximated using a set of weighted samples, $\left\{(s_t^{(i)}, w_t^{(i)}\right\}$, drawn at random from the distribution
  - Samples represent an empirical density function

$$p_N(s) = \sum_{i=1}^{N} w_t^{(i)} \delta_{s_t^{(i)}}(s)$$

  - If the samples are drawn from everywhere in the distribution and if the weight is set appropriately

$$\int_{s_1}^{s_2} p(s)ds \approx \int_{s_1}^{s_2} p_N(s)ds = \sum_{s_t^{(j)} \in [s_1, s_s]} w_t^{(j)}$$

Sampling-Based Filters

- Monte Carlo Sampling from the distribution $p(s)$ produces a sample distribution $p_N(s)$ that approximates $p(s)$ where every sample has a weight of $1/N$
  - Samples ("Particles") can approximately represent any distribution in a finite amount of memory

Sequential Monte Carlo Filters

- Sequential Monte Carlo Filters (Particle filters) are a version of the recursive Bayesian filter that uses samples to represent the distribution
  - Prediction:

$$\{\tilde{s}_t^{(i)}, w_{t-1}^{(i)}\} \quad : \quad \tilde{s}_t^{(i)} \sim p(s_t \mid \tilde{s}_{t-1}^{(i)})$$

  - Measurement:

$$\{\tilde{s}_t^{(i)}, w_t^{(i)}\} \; : \; w_t^{(i)} = \frac{1}{\alpha} w_{t-1}^{(i)} p(o_t \mid \tilde{s}_t^{(i)}), \alpha = \sum_{i=1}^{N} w_{t-1}^{(i)} p(o_t \mid \tilde{s}_t^{(i)})$$

Sequential Monte Carlo Filters

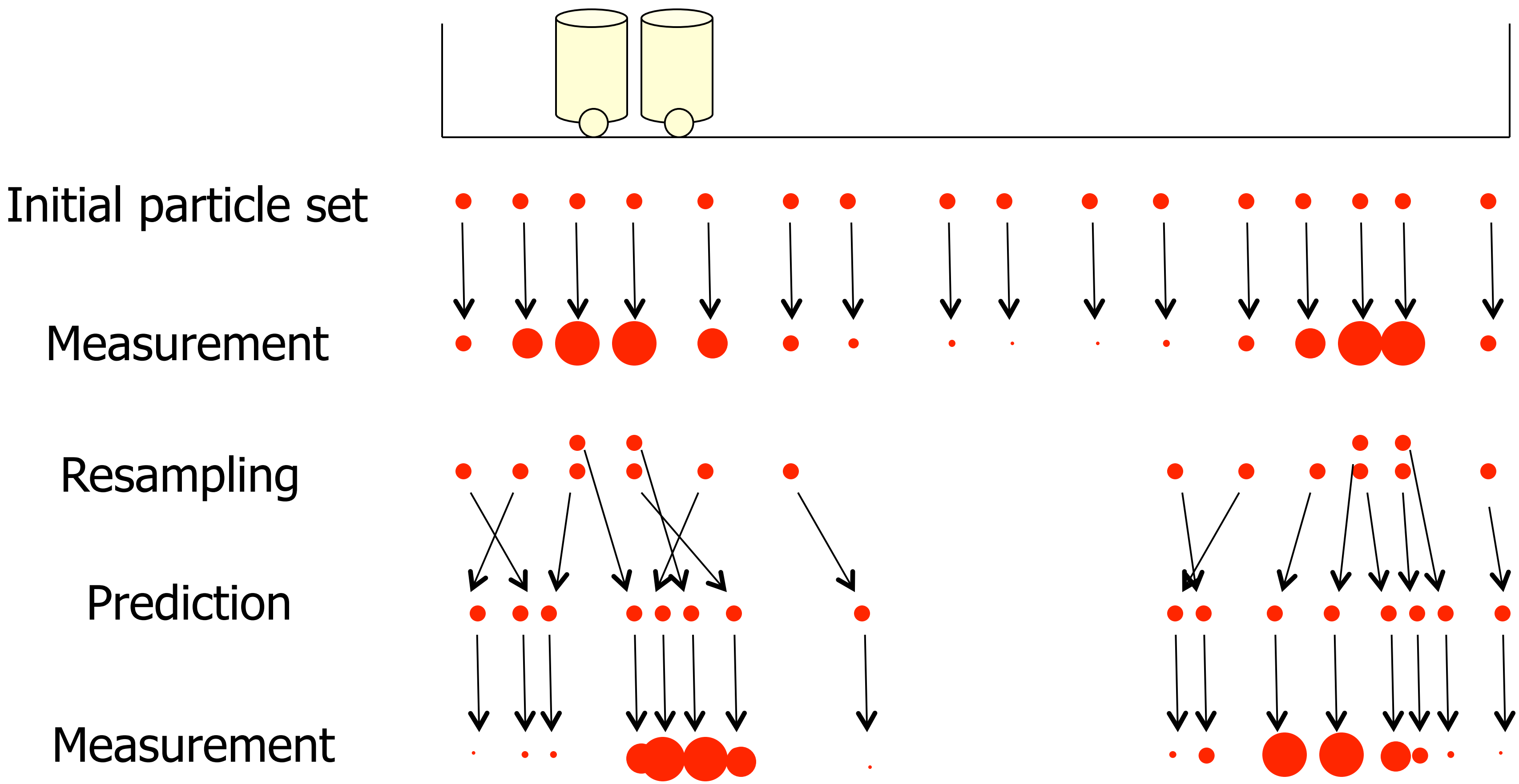
- The basic filter can lead to a degenerate distribution (samples have very uneven weights)
  - A lot of memory might be spent on samples (particles) with weights close to *0*.
  - Loss of quality in the approximation
- Resampling after each iteration

$$\{\widehat{s}_t^{(i)}, \widehat{w}_t^{(i)}\} \quad : \quad \widehat{s}_t^{(i)} \sim w_t^{(i)} \ , \ \widehat{w}_t^{(i)} = \frac{1}{N}$$

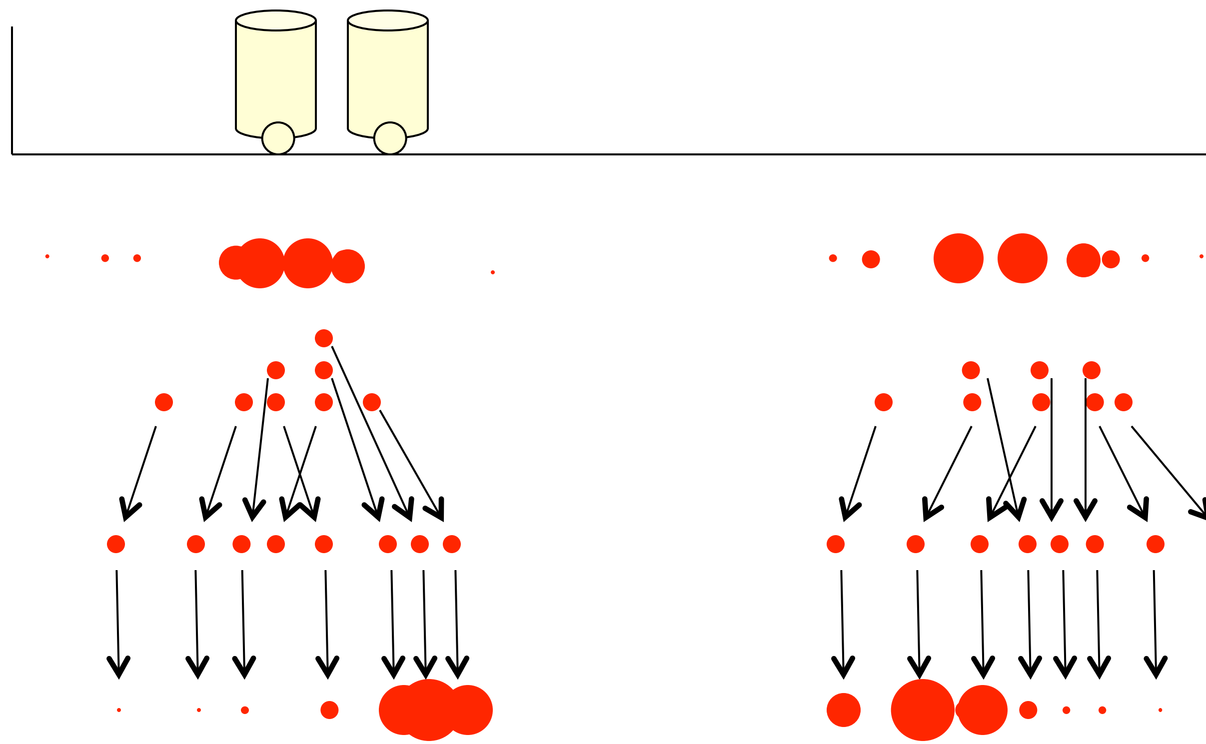
Sequential Monte Carlo Filters

- Simple location estimation problem
  - Robot moves along a hallway, initially not knowing its location or orientation
  - Robot can measure the distance to the closest wall with a noisy omnidirectional sonar sensor

Sequential Monte Carlo Filters



Sequential Monte Carlo Filters



Resampling

Prediction

Measurement

Sequential Monte Carlo Filters

- Particle filters do not impose any limitations on the distributions or process models used
  - Advantages:
    - Arbitrary distributions
    - Arbitrary models
    - Controllable complexity: $O(N)$
  - Disadvantages:
    - Only approximate distribution
    - No obvious estimate (this is a problem with all general distribution estimators)
      - Maximum likelihood ?
      - Minimum squared error ?
      - Highest likelihood region ?

Optimal Estimation

- Different estimators for different problems
  - General Bayesian filter
    - For discrete problems with small state spaces
  - Kalman filters
    - Fast estimators
    - Assumes Gaussian distributions
    - Only suitable for unimodal distributions
  - Discretization
    - For state spaces that form a small number of partitions
    - Only approximate solution
    - Might violate Markov property
  - Particle filters
    - Represents arbitrary processes and distributions
    - Only approximate solution
    - Number of particles (samples) effects precision