Lecture 7: Kalman Filters

CS 395T: Intelligent Robotics
Benjamin Kuipers

Stochastic Models of an Uncertain World

\[ \dot{x} = F(x, u) \quad \dot{x} = F(x, u, \varepsilon_i) \]
\[ y = G(x) \quad y = G(x, \varepsilon_2) \]

- Actions are uncertain.
- Observations are uncertain.
- \( \varepsilon_i \sim N(0, \sigma_i) \) are random variables

Observers

\[ \dot{x} = F(x, u, \varepsilon_i) \]
\[ y = G(x, \varepsilon_2) \]

- The state \( x \) is unobservable.
- The sense vector \( y \) provides noisy information about \( x \).
- An observer \( \dot{x} = \text{Obs}(y) \) is a process that uses sensory history to estimate \( x \).
- Then a control law can be written
  \[ u = H_i(\hat{x}) \]

Kalman Filter: Optimal Observer

(Maybeck’s introduction)

Gaussian (Normal) Distribution

- Completely described by \( N(\mu, \sigma^2) \)
  - Mean \( \mu \)
  - Standard deviation \( \sigma \), variance \( \sigma^2 \)

Estimates and Uncertainty

- Conditional probability density function

\[
\frac{1}{(2\pi\sigma^2)^{D/2}}e^{-\frac{(x_1-\mu_1)^2+\cdots+(x_D-\mu_D)^2}{2\sigma^2}}
\]
The Central Limit Theorem

- The sum of many random variables
  - with the same mean, but
  - with arbitrary conditional density functions,
  converges to a Gaussian density function.

- If a model omits many small unmodeled effects, then the resulting error should converge to a Gaussian density function.

Illustrating the Central Limit Thm

- Add 1, 2, 3, 4 variables from the same distribution.

Estimating a Value

- Suppose there is a constant value \( x \).
  - Distance to wall; angle to wall; etc.
- At time \( t \), observe value \( z \) with variance \( \sigma^2 \)
- The optimal estimate is \( \hat{x}(t) = z \) with variance \( \sigma^2 \)

A Second Observation

- At time \( t \), observe value \( z \) with variance \( \sigma^2 \)

Merged Evidence

- Weighted average of estimates.
  \( \hat{x}(t) = A z_1 + B z_2 \quad A + B = 1 \)
- The weights come from the variances.
  - Smaller variance = more certainty

Update Mean and Variance

- Weighted average of estimates.
  \( \hat{x}(t) = A z_1 + B z_2 \quad A + B = 1 \)
- The weights come from the variances.
  - Smaller variance = more certainty
From Weighted Average to Predictor-Corrector

- Weighted average:
  \[ \hat{x}(t_2) = A\hat{x}_1 + B\hat{x}_2 = (1 - K)\hat{x}_1 + K\hat{x}_2 \]

- Predictor-corrector:
  \[ \hat{x}(t_2) = z_1 + K(z_2 - \hat{x}(t_1)) \]
  \[ \text{This version can be applied “recursively”.} \]

Predictor-Corrector

- Update best estimate given new data
  \[ \hat{x}(t_2) = \hat{x}(t_1) + K(t_2)(z_2 - \hat{x}(t_1)) \]
  \[ K(t_2) = \frac{\sigma^2(t_2)}{\sigma^2(t_2) + \sigma^2(t_1)} \]

- Update variance:
  \[ \sigma^2(t_2) = \sigma^2(t_1) - K(t_2)\sigma^2(t_1) \]
  \[ = (1 - K(t_2))\sigma^2(t_1) \]

Static to Dynamic

- Now suppose \( x \) changes according to
  \[ \dot{x} = F(x, u, \varepsilon) = u + \varepsilon \quad (N(0, \sigma^2)) \]

Dynamic Prediction

- At \( t_2 \) we know \( \hat{x}(t_2) \)
- At \( t_3 \) after the change, before an observation.
  \[ \hat{x}(t_3) = \hat{x}(t_2) + u[t_3 - t_2] \]
  \[ \sigma^2(t_3) = \sigma^2(t_2) + \sigma^2[u] \]

- Next, we correct this prediction with the observation at time \( t_3 \).

Dynamic Correction

- At time \( t_3 \) we observe \( z_3 \) with variance \( \sigma^2(z_3) \)
- Combine prediction with observation.
  \[ \hat{x}(t_3) = \hat{x}(t_3) + K(t_3)(z_3 - \hat{x}(t_3)) \]
  \[ \sigma^2(t_3) = (1 - K(t_3))\sigma^2(t_3) \]
  \[ K(t_3) = \frac{\sigma^2(t_3)}{\sigma^2(t_3) + \sigma^2(t_1)} \]

Qualitative Properties

- Suppose measurement noise \( \sigma^2(z_3) \) is large.
  - Then \( K(t_3) \) approaches 0, and the measurement will be mostly ignored.

- Suppose prediction noise \( \sigma^2(t_1) \) is large.
  - Then \( K(t_3) \) approaches 1, and the measurement will dominate the estimate.
**Kalman Filter**

- Takes a stream of observations, and a dynamical model.
- At each step, a weighted average between
  - prediction from the dynamical model
  - correction from the observation.
- The Kalman gain $K(t)$ is the weighting,
  - based on the variances $\sigma^2(t)$ and $\sigma'$.
- With time, $K(t)$ and $\sigma'(t)$ tend to stabilize.

**Simplifications**

- We have introduced Kalman filters using a one-dimensional system.
  - Most applications are higher dimensional.
- We have assumed the state variable is observable.
  - In general, sense data give indirect evidence.
  $$\dot{x} = F(x,u,\epsilon_1) = u + \epsilon_1$$
  $$z = G(x,\epsilon_2) = x + \epsilon_2$$
- We will discuss the more complex case next.

**Up To Higher Dimensions**

(Former & Bishop version of the story)

- Our previous Kalman Filter discussion was of a simple one-dimensional model.
- Now we go up to higher dimensions:
  - State vector: $\mathbf{x} \in \mathbb{R}^n$
  - Sense vector: $\mathbf{z} \in \mathbb{R}^m$
  - Motor vector: $\mathbf{u} \in \mathbb{R}^l$
- First, a little statistics.

**Expectations**

- Let $x$ be a random variable.
- The expected value $E[x]$ is the mean:
  $$E[x] = \int x \, p(x) \, dx = \frac{1}{N} \sum_{i=1}^{N} x_i$$
  - The probability-weighted mean of all possible values. The sample mean approaches it.
- Expected value of a vector $\mathbf{x}$ is by component.
  $$E[\mathbf{x}] = \bar{x} = [\bar{x}_1, \ldots, \bar{x}_n]^T$$

**Variance and Covariance**

- The variance is $E[(x-E[x])^2]$:
  $$\sigma^2 = E[(x-\bar{x})^2] = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$
- Covariance matrix is $E[(x-E[x])(x-E[x])^T]$:
  $$C_{ij} = \frac{1}{N} \sum_{i=1}^{N} (x_{i_1} - \bar{x}_i)(x_{i_2} - \bar{x}_j)$$
  - Divide by $N-1$ to make the sample variance an *unbiased estimator* for the population variance.

**Biased and Unbiased Estimators**

- Strictly speaking, the sample variance
  $$\sigma^2 = E[(x-\bar{x})^2] = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$
is a biased estimate of the population variance. An unbiased estimator is:
  $$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$
- **But**: “If the difference between $N$ and $N-1$ ever matters to you, then you are probably up to no good anyway…” [Press, et al]
Covariance Matrix
• Along the diagonal, $C_{ii}$ are variances.
• Off-diagonal $C_{ij}$ are essentially correlations.

\[
\begin{bmatrix}
C_{1,1} = \sigma_1^2 & C_{1,2} & \cdots & C_{1,N} \\
C_{2,1} & C_{2,2} = \sigma_2^2 & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
C_{N,1} & \cdots & C_{N,2} & \cdots & C_{N,N} = \sigma_N^2
\end{bmatrix}
\]

Independent Variation
• $x$ and $y$ are Gaussian random variables ($N=100$)
• Generated with $\sigma_x=1$, $\sigma_y=3$
• Covariance matrix:

\[
C_{xy} = \begin{bmatrix}
0.90 & 0.44 \\
0.44 & 8.82
\end{bmatrix}
\]

Dependent Variation
• $c$ and $d$ are random variables.
• Generated with $c=x+y$, $d=x-y$
• Covariance matrix:

\[
C_{cd} = \begin{bmatrix}
10.62 & -7.93 \\
-7.93 & 8.84
\end{bmatrix}
\]

Discrete Kalman Filter
• Estimate the state $x \in \mathbb{R}^n$ of a linear stochastic difference equation
  
  $x_k = Ax_{k-1} + Bu_k + w_k$

  - process noise $w$ is drawn from $N(0, Q)$, with covariance matrix $Q$.
• with a measurement $z \in \mathbb{R}^m$
  
  $z_t = Hx_t + v_t$

  - measurement noise $v$ is drawn from $N(0, R)$, with covariance matrix $R$.
• $A$, $Q$ are $n \times n$. $B$ is $n \times l$. $R$ is $m \times m$. $H$ is $m \times n$.

Estimates and Errors
• $\hat{x}_k \in \mathbb{R}^n$ is the estimated state at time-step $k$.
• $\hat{x}_k \in \mathbb{R}^n$ after prediction, before observation.
• Errors: $e_k = x_k - \hat{x}_k$
• Error covariance matrices:
  
  $P_k = E[e_k e_k^T]$
  
  $P_{\hat{x}} = E[\hat{x}_k e_k^T]$

• Kalman Filter’s task is to update $\hat{x}_k$, $P_k$

Time Update (Predictor)
• Update expected value of $x$
  
  $\hat{x}_k = A\hat{x}_{k-1} + Bu_k$

• Update error covariance matrix $P$
  
  $P_{\hat{x}} = AP_{\hat{x}}A^T + Q$

• Previous statements were simplified versions of the same idea:
  
  $\hat{x}(t_3) = \hat{x}(t_2) + u[t_3 - t_2]$
  
  $\sigma^2(t_3) = \sigma^2(t_2) + \sigma^2[t_3 - t_2]$
Measurement Update (Corrector)

- Update expected value
  \[ \hat{x}_k = \hat{x}_k + K_k (z_k - H\hat{x}_k) \]
  \[ \text{– innovation is } \quad z_k - H\hat{x}_k \]
- Update error covariance matrix
  \[ P_k = (I - K_kH)P_k^* \]

  \[ \text{Compare with previous form} \]
  \[ \hat{x}(t_3) = \hat{x}(t_3) + K(t_3)(z_3 - \hat{x}(t_3)) \]
  \[ \sigma^2(t_3) = (1 - K(t_3))\sigma^2(t_3) \]

The Kalman Gain

- The optimal Kalman gain \( K_k \) is
  \[ K_k = P_k^*H^T(HP_kH^T + R)^{-1} \]
  \[ = \frac{P_k^*H^T}{HP_kH^T + R} \]

  \[ \text{Compare with previous form} \]
  \[ K(t_3) = \frac{\sigma^2(t_3)}{\sigma^2(t_3) + \sigma^2} \]

Extended Kalman Filter

- Suppose the state-evolution and measurement equations are non-linear:
  \[ x_k = f(x_{k-1}, u_k) + w_{k-1} \]
  \[ z_k = h(x_k) + v_k \]
  \[ - \text{process noise } w \text{ is drawn from } N(0,Q), \text{ with covariance matrix } Q. \]
  \[ - \text{measurement noise } v \text{ is drawn from } N(0,R), \text{ with covariance matrix } R. \]

The Jacobian Matrix

- For a scalar function \( y = f(x) \),
  \[ \Delta y = f'(x)\Delta x \]
- For a vector function \( y = f(x) \),
  \[ \Delta y = J\Delta x = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(x)}{\partial x_n} & \cdots & \frac{\partial f_n(x)}{\partial x_n} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{bmatrix} \]

Linearize the Non-Linear

- Let \( A \) be the Jacobian of \( f \) with respect to \( x \).
  \[ A_j = \frac{df}{dx_j}(x_{k-1},u_k) \]
- Let \( H \) be the Jacobian of \( h \) with respect to \( x \).
  \[ H_j = \frac{dh}{dx_j}(x_k) \]
- Then the Kalman Filter equations are almost the same as before!

EKF Update Equations

- Predictor step:
  \[ \hat{x}_k = f(\hat{x}_{k-1},u_k) \]
  \[ P_k^* = AP_kA^T + Q \]
- Kalman gain:
  \[ K_k = P_k^*H^T(HP_k^*H^T + R)^{-1} \]
- Corrector step:
  \[ \hat{x}_k = \hat{x}_k + K_k(z_k - h(\hat{x}_k)) \]
  \[ P_k = (I - K_kH)P_k^* \]
What Have We Got?

- If we have a stochastic dynamical model of how a system behaves, and
- If we have a stream of data about the system, and
- If we can represent uncertainty (in the model and in the data) with Gaussian distributions, then
- We can merge data with predictions in a good (sometimes optimal) way.

TTD

- Intuitive explanations for $\text{APA}^T$ and $\text{HPH}^T$ in the update equations.