LECTURE 2

* Kalman Filter

The recursive LS formulation from last time was:

\[ P_{i-1} = P_{i-1}^{-1} A_{i}^{T} R_{i}^{-1} A_{i} \]

\[ K_{i} = P_{i} A_{i}^{T} R_{i}^{-1} \]

\[ \hat{X}_{i} = \hat{X}_{i-1} + k_{i} (Y_{i} - A_{i} \hat{X}_{i-1}) \]

\[ \Rightarrow \quad \text{The fundamental assumption is} \]

that \( X \) is constant and that \( \hat{X} \) must be

updated as new data points arrive.

Let's assume now that \( X \) can actually evolve in time:

\[ X_{i} = F_{i-1} X_{i-1} + \epsilon_{i-1} \]

\[ X_{i} = F_{i-1} X_{i-1} + G_{i-1} u_{i-1} + \epsilon_{i-1} \]

or

\[ Y_{i} = A_{i} X_{i} + \eta_{i} \]

\[ Y_{i} = A_{i} X_{i} + \eta_{i} \]

\( \epsilon_{i} \) = noise in the model

\( \eta_{i} \) = noise in the measurements \( \Rightarrow \) it's the "&" in lecture 1

\( u_{i} \) = control vector (if any)

Let's also assume:

\( \epsilon_{i} \sim N(0, Q_{i}) \) \( \Rightarrow \) Gaussian

and \( F_{i}, G_{i}, A_{i} \) known

\( \eta_{i} \sim N(0, R_{i}) \) \( \Rightarrow \) noise for every \( i \)

Because of \( \epsilon_{i}, \eta_{i} \), \( X_{i} \) is a random variable \( \Rightarrow \) Our estimation will

be on the mean and we can write:

\[ E(X_{i}) = E \left( F_{i-1} X_{i-1} + G_{i-1} u_{i-1} + \epsilon_{i-1} \right) \]

\[ = F_{i-1} E(X_{i-1}) + G_{i-1} u_{i-1} \]

\[ \Rightarrow \text{The mean of } X_{i} \text{ is propagated by the formula:} \]
\[ \bar{x}_i = F_{i-1} \bar{x}_{i-1} + G_{i-1} u_{i-1} \quad \text{where} \quad \bar{x}_i \triangleq E(X_i) \quad (1) \]

\[
E[(x_i - \bar{x}_i)(x_i - \bar{x}_i)^T] = \\
= E\left[ (F_{i-1} x_{i-1} + G_{i-1} u_{i-1} + \varepsilon_{i-1} - F_{i-1} \bar{x}_{i-1} - G_{i-1} u_{i-1}) (\ldots) ^T \right] \\
= E\left[ (F_{i-1} (x_{i-1} - \bar{x}_{i-1}) + \varepsilon_{i-1}) (F_{i-1} (x_{i-1} - \bar{x}_{i-1}) + \varepsilon_{i-1}) ^T \right] \\
= F_{i-1} E\left[ (x_{i-1} - \bar{x}_{i-1})(x_{i-1} - \bar{x}_{i-1})^T \right] F_{i-1}^T + E\left[ \varepsilon_{i-1} \varepsilon_{i-1}^T \right] \\
= 0 \\
= 0
\]

By definition, the noise is uncorrelated with the variable \( X_i \).

Because \( E[(x_i - \bar{x}_i)(x_i - \bar{x}_i)^T] = P_i \) - covariance of the estimation, we can write:

\[ P_i = F_{i-1} P_{i-1} F_{i-1}^T + Q_{i-1} \quad (2) \]

Note: In eq. (1) and (2), we have just used the evolution model of \( X_i \); not the estimations \( \hat{x}_i, \hat{P}_i \). When it comes about estimation and we are at time \( i \), two options are viable:
\[ \hat{X}_i^- \quad \hat{P}_i^- \quad \hat{X}_i^+ \quad \hat{P}_i^+ \]

a) We use data collected up to \( i \), EXCLUDED
   - the data collected at time \( i \) \( \Rightarrow \) Let's use
   - the notation: \( \hat{X}_i^- \quad \hat{P}_i^- \)

b) We use data collected up to \( i \), INCLUDED
   - the data collected at time \( i \) \( \Rightarrow \) Let's use
   - the notation: \( \hat{X}_i^+ \quad \hat{P}_i^+ \)

Therefore, we can have a two-tiered approach:

- If we are waiting for the data to be collected at time \( i \), we can
  enhance our estimation from time \( i-1 \) by using the model:

\[
\begin{align*}
\hat{X}_i^- &= F_{i-1} \hat{X}_{i-1}^- + G_{i-1} u_{i-1} \\
\hat{P}_i^- &= F_{i-1} \hat{P}_{i-1}^- F_{i-1}^T + Q_{i-1}
\end{align*}
\]

\((*)\)

- If we have collected the data at time \( i \), we can refine the estimation
  by using the recursive Least square formula:

\[
\hat{P}_i^+ = \left[ (\hat{P}_i^-)^{-1} + A_i^T R_i^{-1} A_i \right]^{-1}
\]

\((***)\)

\[
\begin{align*}
\kappa_i &= \hat{P}_i^+ A_i^T R_i^{-1} \\
\hat{X}_i^+ &= \hat{X}_i^- + \kappa_i (Y_i - A_i \hat{X}_i^-)
\end{align*}
\]

\((*)\) is a prediction step and \((***)\) is a correction step \( \Rightarrow \) The
combination is the Kalman filter

Note a few things:

\[
\begin{align*}
\hat{X}_i^+ &= F_{i-1} \hat{X}_{i-1}^+ + G_{i-1} u_{i-1} + K_i Y_i - K_i A_i F_{i-1} \hat{X}_{i-1}^+ - K_i A_i G_{i-1} u_{i-1}
\end{align*}
\]
\[
(\hat{X}_{i-1} - K_i A_i \hat{X}_{i-1})_{X_i-1} + (G_{i-1} - K_i A_i G_{i-1}) u_{i-1} + K_i Y_i
\]

\[kF\] is actually a time-varying filter that operates on the input variables \(u\) and \(Y\).

In (*), we provide an estimation for \(X\) in the absence of new data. If no further data were provided, our estimation would evolve simply based on the model. Uncertainty would grow in time because of the term \(P_{i-1}\). Estimation worsens between measures.

In (**), a large \(P_i\) is overshadowed by \(A_i^T P_{i-1} A_i\), since it contributes as \((\hat{P}_i^{-1})^{-1}\), so a small uncertainty on the measurements (i.e., \(R_i\) small) will likely make \(\hat{P}_i +\) small, i.e.,

\[P_i + = (A_i^T P_{i-1} A_i)^{-1}\]

However, in \(K_i = P_i + A_i^T P_{i-1}\), a small \(P_i +\) is bad for updating the estimation \(\hat{X}_i +\). That's why the presence of \(P_{i-1}\) compensates for \(P_i +\) and makes \(K_i\) sizable enough to use \(Y_i\) in the update of \(\hat{X}_i +\).

Moreover, note:\n- \(R_i\) large ⇒ \(A_i^T R_{i-1} A_i\) small ⇒ \(P_i + \approx P_i -\) and \(K_i\) small ⇒ The uncertainty of the estimation still evolves as in the absence of new measurements and new measurements are trusted very little when it comes about updating the estimate of \(X\).

Finally note: while a small \(R_i\) is generally useful (i.e., we can trust the measurements), a small \(Q_i\) may eventually lead to undesired conditions:
- \(Q_i\) small ⇒ \(P_i\) small ⇒ \(P_i + \approx P_i -\) small ⇒ \(K_i\) small, i.e., if we
trust the model too much, the measurements will be eventually ignored.

* Kalman Filter for Nonlinear Models

\[ X_i = F_{i-1} X_{i-1} + G_{i-1} u_{i-1} + e_{i-1} \]
\[ Y_i = A_i X_i + \eta_i \]

If the model is accurate, then the error \( Y_i - A_i \hat{X}_i \) → Gaussian RV

The KF is able to extract all the available information buried in noisy data

What if the model is nonlinear?

\[ X_i = f_i(X_{i-1}, u_{i-1}) + e_{i-1} \]
\[ Y_i = g_i(X_i) + \eta_i \]

\[ e_i \sim N(0, Q_i) \]
\[ \eta_i \sim N(0, R_i) \]

EKF: One approach focuses on the variations of \( X_i \) from its expected value \( \mu_i \) and uses the first order Taylor series expansion:

\[ f_i(X_{i-1}, u_{i-1}) \approx f_i(\mu_{i-1}, u_{i-1}) + \frac{df_i}{dX}(\mu_{i-1}, u_{i-1})(X_i - \mu_{i-1}) \]

\[ g_i(X_i) \approx g_i(\mu_i) + \frac{dg_i}{dX}(\mu_i)(X_i - \mu_i) \]

\[ \mu_i = E(X_i) = E(f_i(X_{i-1}, u_{i-1})) \approx f_i(\mu_{i-1}, u_{i-1}) \]

Define:

\[ Z_i = X_i - \mu_i \]
\[ V_i = Y_i - g_i(\mu_i) \]
\[ F_{i-1} = \frac{df_i}{dX}(\mu_{i-1}, u_{i-1}) \]
\[ A_i = \frac{dg_i}{dX}(\mu_i) \]

and consider the linearized model:

\[ Z_i = F_{i-1} Z_{i-1} + E_{i-1} \]
\[ V_i = A_i Z_i + \eta_i \]

(1)
Now, in order to estimate $X_i$, one can observe:

Our estimation focuses on the expected value of $X_i$ (so it was in the linear case) \(\Rightarrow\) Denoted with $\hat{X}_{i-1}^+$ the best estimate of $\mu_{i-1}$ after collecting the data at time $i-1$, we can write:

\[
\hat{X}_{i-1}^- = f_i(\hat{X}_{i-1}^+, u_{i-1})
\]

\[
F_{i-1} = \frac{d}{dX} f_i(\hat{X}_{i-1}^+, u_{i-1})
\]

Also note: $P_i \triangleq E((X_i - \mu_i)(X_i - \mu_i)^T) = E(Z_i Z_i^T) \Rightarrow$ We can apply the KF to (1) and estimate $P_i$:

\[
\hat{P}_{i-1}^- = F_{i-1}^- P_{i-1} F_{i-1}^T + Q_{i-1}
\]

\(\hat{X}_{i-1}^- = f_i(\hat{X}_{i-1}^+, u_{i-1})\)

\(\hat{F}_{i-1} = \frac{d}{dX} f_i(\hat{X}_{i-1}^+, u_{i-1})\)

Similarly note:

\[
\begin{align*}
\hat{Z}_{i}^+ &= \hat{Z}_{i}^- + k_i (Y_i - A_i \hat{Z}_{i}^-) \\
\hat{X}_{i}^+ &= \hat{X}_{i}^- + k_i (Y_i - g_i(\mu_i) - A_i (\hat{X}_{i}^- - \mu_i)) \\
&= g_i (\hat{X}_{i}^-) \\
&\Rightarrow \hat{X}_{i}^+ = \hat{X}_{i}^- + k_i (Y_i - g_i(\hat{X}_{i}^-))
\end{align*}
\]

Remember: we need to update the estimate of $P_i$ in order to compute $K_i$:

\[
\hat{P}_{i-1}^+ = E((\hat{X}_{i}^+-\mu_i)(\hat{X}_{i}^+-\mu_i)^T) = E(\hat{Z}_{i}^+ \hat{Z}_{i}^{+-T}) \Rightarrow \text{By using KF on (1):}
\]
\[
\hat{P}_{i+} = \left[ (\hat{P}_i)^{-1} + A_i^T R_i^{-1} A_i \right]^{-1}
\]

where:
- \( \hat{X}_i^- \) is the best estimate of \( X_i \) before collecting the data at time \( i \)
- \( A_i = \frac{\partial g_i}{\partial \hat{X}_i} (\hat{X}_i^-) \)

The combination (E1)-(E2) is called "Extended Kalman Filter" (EKF)

Problems:

- \( F_{i-1} \) and \( A_i \) are obtained by linearizing around the estimated \( \hat{X}_{i-1} \) and \( \hat{X}_i^- \), respectively, not the actual mean values \( \mu_i, \mu_i^- \), respectively ⇒ If the estimates are bad, the filter may drift away

    \[
    \text{EKF can give acceptable results as long as the nonlinearity is weak (i.e., the linearization introduces a small error)}
    \]

- \( F_{i-1} \) and \( A_i \) are Jacobian matrices ⇒ Their calculation can be challenging and numerical issues can rise.

- By linearizing, we assume that the distribution of \( X \) propagates from time \( i-1 \) to time \( i \) according to a linear model ⇒ It is not always the case:
Now, when we propagate our estimation by using KF (linear case) and EKF (nonlinear case), we have:

- **KF**
  - Linear Model
  - The probability distribution of $\hat{X}_i$ is approximated with a Gaussian one.

- **EKF**
  - Nonlinear Model
  - This is the expected value of the distribution.
  - With EKF, $\hat{X}_i$ may be a very bad estimation of $\mu_i \Rightarrow$ We need a different approach that aims to sample the distributions of $\hat{X}_i$, $\hat{X}_i^+$ rather than just assume they are Gaussian.

**UKF**

- First, let's assume that $\hat{X}_i^+$ is Gaussian and let's propagate many values drawn from this distribution through the nonlinear model.

- Second, as the number of values (a.k.a. "particles") increases, a better estimation of the distribution of $\hat{X}_i$ can be given.

We don't need to follow a Monte Carlo approach, though. It can be proved that - if $\hat{X}_i^+ \text{ is Gaussian}$ and $n$ is the size of $X$ - we only need $n+1$ particles.
to estimate mean and variance of distribution determined by the nonlinear transformation of $\hat{X}_{i-1}$ we have:

$\hat{X}_{i-1}$ is the mean of the input distribution

$\hat{P}_{i-1}$ is the covariance matrix of the input

$\Rightarrow$ The $n+1$ points must be:

distribution $\Rightarrow$ It is a $n \times n$ matrix with columns:

$\begin{bmatrix}
\hat{p}_{i-1,1}^+ & \hat{p}_{i-1,2}^+ & \cdots & \hat{p}_{i-1,n}^+
\end{bmatrix}$

$\sigma_0 = \hat{X}_{i-1}$

$\sigma_j = \sigma_0 + \sqrt{n} \frac{\hat{p}_{i-1,j}^+}{j = 1, 2, \ldots, n}$

$\sigma_{n+j} = \sigma_0 - \sqrt{n} \frac{\hat{p}_{i-1,j}^+}{j = 1, 2, \ldots, n}$

$\Rightarrow$ The transformation of these $2n+1$ points is:

$\gamma_0 = f_i(\sigma_0, u; i-1)$

$\gamma_j = f_i(\sigma_j, u; i-1), j = 1, 2, \ldots, 2n$

Sample mean and covariance of this distribution can be used for prediction:

$\hat{X}_i = \frac{1}{2n+1} \sum_{j=0}^{2n} \gamma_j$

$\hat{P}_i = \text{cov}(\sum_{j=0}^{2n} \gamma_j)(\sum_{j=0}^{2n} \gamma_j)^T + \Omega_{i-1}$

$= \frac{1}{2n+1} \sum_{j=0}^{2n} (\gamma_j - \hat{X}_i)(\gamma_j - \hat{X}_i)^T + \Omega_{i-1}$

Analogously, one can compute:

$\xi_0 = g_i(\sigma_0)$

$\Rightarrow$ One can estimate sample mean and covariance:

$\hat{Y}_i = \frac{1}{2n+1} \sum_{j=0}^{2n} \xi_j$

$P_{yy} = \frac{1}{2n+1} \sum_{j=0}^{2n} (\xi_j - \hat{Y}_i)(\xi_j - \hat{Y}_i)^T + \Omega_i$
Note that we add $Q_i$ and $Q_i$ to $\hat{P}_i^-$ and $\hat{P}_i$, respectively, because we want to compute an estimation of the covariances of samples of $X$ and $Y$, respectively, but the samples have been chosen in a deterministic way.

With $\sum_x x_j \tilde{x}$ and $\sum_y y_j \tilde{y}$, one can also compute:

$$p_{xy} = \frac{1}{2n+1} \sum_{j=0}^{2n} (x_j - \bar{x})(y_j - \bar{y})$$

If $X$ and $Y$ have size $n$ and $m$, respectively, this is a $n \times m$ matrix.

It can be shown that the value:

$$\hat{x}_i^+ = \hat{x}_i^- + k_i (Y_i - \bar{Y}_i) \quad \text{with} \quad k_i = p_{xy} p_{yy}^{-1}$$

is the unbiased estimation of $X_i$ that minimizes the sum of the variances of the individual components of $X_i$. Therefore, the following filter is derived:

**Sampling Step:**

$$\sigma_0 = \hat{x}_{i-1}^+ \quad ; \quad \sigma_j = \hat{x}_{i-1}^+ + \sqrt{n} \sum_{j=0}^{2n} \hat{p}_{i-1}^+ \quad ; \quad \sigma_{n+j} = \hat{x}_{i-1}^+ - \sqrt{n} \sum_{j=0}^{2n} \hat{p}_{i-1}^+ \quad j = 1, 2, \ldots, n$$

(1)

with $[\hat{p}_{i-1}^+]_j = \hat{p}_{i+1}^+ \quad j$-th column of $\hat{P}_{i-1}^+$

$$\gamma_j = \tilde{y}_j (\sigma_j X_{i-1}) \quad j = 0, 1, \ldots, 2n$$

$$\bar{\gamma}_j = \frac{\gamma_j}{\sigma_j} \quad j = 0, 1, \ldots, 2n$$

**Prediction Step:**

$$\hat{x}_i^- = \frac{1}{2n+1} \sum_{j=0}^{2n} \gamma_j \quad \hat{P}_i^- = \frac{1}{2n+1} \sum_{j=0}^{2n} (x_j - \hat{x}_i^-)(x_j - \hat{x}_i^-)^\top + Q_{i-1}$$
Correction Step:

\[(U3) \quad \tilde{y}_i = \frac{1}{2n+1} \sum_{j=0}^{2n} x_j \quad P_{yy} = \frac{1}{2n+1} \sum_{j=0}^{2n} (x_j - \tilde{y}_i) (x_j - \tilde{y}_i)^T + R_i \]

\[P_{xy} = \frac{1}{2n+1} \sum_{j=0}^{2n} (x_j - \tilde{x}_i^-) (x_j - \tilde{y}_i) \]

\[\hat{P}_i^+ = \hat{P}_i^- - P_{xy} P_{yy}^{-1} P_{xy}^T \]

\[\hat{x}_i^+ = \hat{x}_i^- + P_{xy} P_{yy}^{-1} (y_i - \tilde{y}_i) \]

The combination of steps (U1) - (U3) is the "Unscented Kalman Filter" (UKF).

Slides present examples of application of Kalman filter to neural decoding of kinematic variables.