CS294-40 Learning for Robotics and Control Lecture 14 - 10/14/2008

Kalman Filtering, EKF, Unscented KF, Smoother, EM

Lecturer: Pieter Abbeel *Scribe: Jared Wood*

1 Kalman Filtering Recap

Recall the linear system

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_t + v_t \end{aligned} \tag{1}$$

where $w_t \sim N\left(0, \Sigma_w\right)$, $v_t \sim N\left(0, \Sigma_v\right)$, and $x_0 \sim N\left(x_{0|-1}, P_{0|-1}\right)$. Note that $x \sim N\left(\mu, \Sigma\right)$ means

$$P\left(x\right) = \frac{1}{\left(2\pi\right)\left|\Sigma\right|^{1/2}} e^{-\frac{1}{2}\left(x-\mu\right)^{\mathsf{T}}\Sigma^{-1}\left(x-\mu\right)}.$$

We also have: $Ex = \mu$ and $E\left(x-\mu\right)\left(x-\mu\right)^{\mathsf{T}} = \Sigma$.

We will use the following notation:

$$\begin{aligned} \hat{x}_{t|t} &= E\left[x_t | y_{0:\,t}\right] \\ P_{t|t} &= E\left[\left(x_t - \hat{x}_{t|t}\right)\left(x_t - \hat{x}_{t|t}\right)^{\mathsf{T}} | y_{0:\,t}\right] \\ \hat{x}_{t+1|t} &= E\left[x_{t+1} | y_{0:\,t}\right] \\ P_{t+1|t} &= E\left[\left(x_{t+1} - \hat{x}_{t+1|t}\right)\left(x_{t+1} - \hat{x}_{t+1|t}\right)^{\mathsf{T}} | y_{0:\,t}\right] \end{aligned}$$

Note that because $x_{t|\cdot}$ is a Gaussian random variable, it is sufficient to only keep track of the mean and covariance. We can do so by the following computations at each time t:

$$\begin{aligned} \hat{x}_{t+1|t} &= A\hat{x}_{t|t} + Bu_t \\ P_{t+1|t} &= AP_{t|t}A^{\mathsf{T}} + \Sigma_w \\ \hat{x}_{t+1|t+1} &= \hat{x}_{t+1|t} + K_{t+1}\left(y_{t+1} - C\hat{x}_{t+1|t}\right) \\ K_{t+1} &= P_{t+1|t}C^{\mathsf{T}}\left(CP_{t+1|t}C^{\mathsf{T}} + \Sigma_v\right)^{-1} \\ P_{t+1|t+1} &= P_{t+1|t} - P_{t+1|t}C^{\mathsf{T}}\left(CP_{t+1|t}C^{\mathsf{T}} + \Sigma_v\right)^{-1}CP_{t+1|t} \end{aligned} \tag{2}$$

2 Log Likelihood of the Observations

In practice, we only observe the vector $y_{0:T}$. The state estimates we obtain heavily depend on our choice of the covariance matrices Σ_w, Σ_v. How can we decide on a good choice for the covariance matrices?

We will rely upon a very common formalism in statistics: we will estimate the unknown parameters (covariances) by maximizing the (log) likelihood of the observed data. I.e., we find the covariances by solving the following optimization problem:

$$\max_{\Sigma_w, \Sigma_v} ll(\Sigma_v, \Sigma_w) = \max_{\Sigma_w, \Sigma_v} \log P(y_{0:T}; \Sigma_v, \Sigma_w).$$

First, we describe how to evaluate the log-likelihood.

Recall, we have the following for the distribution of y_{t+1} given the observations $y_0, \ldots, y_t$:

$$P\left(y_{t+1}|y_{0:\,t}\right) = \frac{1}{\left(2\pi\right)^{d/2}|\Sigma_{y_{t+1|t}}|^{1/2}} e^{-\frac{1}{2}\left(y_{t+1}-\hat{y}_{t+1|t}\right)^{\mathsf{T}}\Sigma_{y_{t+1|t}}\left(y_{t+1}-\hat{y}_{t+1|t}\right)} \tag{3}$$

where

$$\begin{aligned} \hat{y}_{t+1|t} &= C\hat{x}_{t+1|t} \\ \Sigma_{y_{t+1|t}} &= CP_{t+1|t}C^{\mathsf{T}} + \Sigma_v. \end{aligned} \tag{4}$$

Using Bayes' rule, we write the likelihood of $y_{0:T}$ as

$$\begin{aligned} P\left(y_0, ..., y_T\right) &= P\left(y_0\right)\prod_{t=1}^{T} P\left(y_t|y_{0:t-1}\right) \\ &= \prod_{t=1}^{T}\frac{1}{\left(2\pi\right)^{d/2}|\Sigma_{y_{t+1|t}}|^{1/2}} e^{-\frac{1}{2}\left(y_{t+1}-\hat{y}_{t+1|t}\right)^{\mathsf{T}}\Sigma^{-1}_{y_{t+1|t}}\left(y_{t+1}-\hat{y}_{t+1|t}\right)} \end{aligned}$$

Now by augmenting the Kalman filter with

$$\begin{aligned} &ll = 0 \\ &ll+ = \log P\left(y_{t+1}|y_{0:\,t}\right) \end{aligned}$$

we can find $\log P\left(y_0, ..., y_T\right)$.

Hence we can efficiently compute the log-likelihood by adding a minor computation to the Kalman filter updates. A simple method for finding the covariance matrices Σ_w, Σ_v that maximize $ll(\Sigma_w, \Sigma_v)$ is to numerically compute the gradient and perform gradient ascent. This works reasonably well for a small number of parameters, yet below we will describe an EM algorithm that tends to work better in practice.

3 Kalman Smoother

So far we have only considered $P\left(x_t|y_{0:\,t}\right)$. I.e., the filtered estimate of x_t only takes into account the "past" information relative to x_t. By incorporating the "future" observations relative to x_t, we can obtain a more refined state estimate.

Estimators that take into account both past and future are often called "smoothers." The Kalman smoother estimates $P\left(x_t|y_{0:\,T}\right)$.

Without a derivation, we state the Kalman smoother equations here:

$$\hat{x}_{t|T} = \hat{x}_{t|t} + L_t\left(x_{t+1|T} - \hat{x}_{t+1|T}\right) \tag{5}$$

$$P_{t|T} = P_{t|t} + L_t\left(P_{t+1|T} - P_{t+1|t}\right)L_t^{\mathsf{T}} \tag{6}$$

$$L_t = P_{t|t}A^{\mathsf{T}}P^{-1}_{t+1|t} \tag{7}$$

Before running the smoother, we must first run the filter. The smoother then proceeds backward in time.

Note that $P_{t+1|T} - P_{t+1|t} < 0$ as the uncertainty over x_{t+1} is smaller when conditioned on all observations, than when only conditioned on past observations.

4 EM Algorithm

The EM algorithm is an efficient way to find the parameters that maximize the log-likelihood.

Without derivation, we provide the algorithm:

- Initialize Σ_w, Σ_v.
- Iterate:
  - Run Kalman filter.
  - Run Kalman smoother.
  - Update Σ_w, Σ_v.

Here, Σ_w and Σ_v are updated as follows:

$$\Sigma_w = \frac{1}{T}\sum_{t=0}^{T-1}\left(\hat{x}_{t+1|T} - A\hat{x}_{t|T} - Bu_t\right)\left(\hat{x}_{t+1|T} - A\hat{x}_{t|T} - Bu_t\right)^{\mathsf{T}} + A_t P_{t|T} A_t^{\mathsf{T}} + P_{t+1|T} - P_{t+1|T} L_t^{\mathsf{T}} A^{\mathsf{T}} - A L_t P_{t+1|T}$$

$$\Sigma_v = \frac{1}{T+1}\sum_{t=0}^{T}\left(y_t - C\hat{x}_{t|T}\right)\left(y_t - C\hat{x}_{t|T}\right)^{\mathsf{T}} + C P_{t|T} C^{\mathsf{T}}$$

5 Extended Kalman Filter

Now consider a nonlinear extension to the Kalman filter. Now the system is

$$\begin{aligned} x_{t+1} &= f\left(x_t, u_t\right) + w_t \\ y_t &= h\left(x_t\right) + v_t \end{aligned} \tag{8}$$

Form an augmented system as

$$\begin{aligned} \begin{pmatrix} x_{t+1} \\ 1 \end{pmatrix} &= A_t \begin{pmatrix} x_t \\ 1 \end{pmatrix} + \begin{pmatrix} B_t \\ 0 \end{pmatrix} + \begin{pmatrix} w_t \\ 0 \end{pmatrix} \\ y_t &= C_t \begin{pmatrix} x_t \\ 1 \end{pmatrix} + \begin{pmatrix} v_t \\ 0 \end{pmatrix} \end{aligned} \tag{9}$$

where

$$\begin{aligned} A_t &= \begin{bmatrix} \frac{\partial f}{\partial x}|_{x=\hat{x}_{t|t}, u=u_t} & f\left(\hat{x}_{t|t}, u_t\right) - \frac{\partial f}{\partial x}|_{x=\hat{x}_{t|t}, u=u_t}\hat{x}_{t|t} \\ 0 & 1 \end{bmatrix} \\ C_t &= \begin{bmatrix} \frac{\partial h}{\partial x}|_{x=\hat{x}_{t|t-1}} & h\left(\hat{x}_{t|t-1}\right) - \frac{\partial h}{\partial x}|_{x=\hat{x}_{t|t-1}}\hat{x}_{t|t} \end{bmatrix} \end{aligned}$$

6 Unscented Kalman Filter

High-level idea: Sample $x_t^{(0)}, ..., x_t^{(m)}$ from $P(x_t)$. Then compute transition according to $f(x_t)$, which results in $(x_{t+1}^{(0)}, ..., x_{t+1}^{(m)})$. You could fit Gaussian to these sample distribution over the state x_{t+1} using weights as

$$\begin{aligned} \hat{x}_{t+1|t} &= \sum_i w_i x_{t+1}^{(i)} \\ P_{t+1|t} &= \sum_i w_i (x_{t+1}^{(i)} - \hat{x}_{t+1|t})(x_{t+1}^{(i)} - \hat{x}_{t+1|t})^{\mathsf{T}} \end{aligned} \tag{10}$$

where $\sum_i w_i = 1$.

Now, with noise, $x_{t+1} = f(x_t, u_t, w_t)$. Sample from

$$\begin{bmatrix} x_t \\ w_t \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{x}_{t+1|t} \\ 0 \end{bmatrix}, \begin{bmatrix} P_{t+1|t} & 0 \\ 0 & \Sigma_w \end{bmatrix}\right)$$

which gives $(x_t^{(0)}, w_t^{(0)}), ..., (x_t^{(m)}, w_t^{(m)})$. Then compute the transition by passing each of these samples through f. $\hat{x}_{t+1|t}$ and $P_{t+1|t}$ are then given by empirical estimates.

The measurements are $y_t = h(x_t, v_t)$. So sample $(x_t^{(0)}, v_t^{(0)}), ..., (x_t^{(m)}, v_t^{(m)})$ from

$$\begin{bmatrix} x_t \\ v_t \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{x}_{t|t-1} \\ 0 \end{bmatrix}, \begin{bmatrix} P_{t|t-1} & 0 \\ 0 & \Sigma_v \end{bmatrix} \right)$$

Then obtain $y_t^{(i)}$ by passing samples through h. Then

$$\begin{bmatrix} x_t \\ v_t \\ y_t \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{x}_{t|t-1} \\ 0 \\ \hat{y}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \hat{\Sigma}_{xx} & 0 & \hat{\Sigma}_{xy} \\ 0 & \hat{\Sigma}_{vv} & \hat{\Sigma}_{vx} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_{yv} & \hat{\Sigma}_{yy} \end{bmatrix} \right)$$

Note that linearization was not necessary. However, many samples were required. This leads to the question of whether the samples can be chosen wisely and thus much fewer samples are required. This is the concept of the sigma point filter. You pick $2L + 1$ points and weights for an L dimensional distribution. For example, a two dimensional Gaussian would require 5 points (one being the mean). The points would be chosen to preserve the covariance and mean of the distribution. If f is linear or quadratic, this sampling yields the exact moment of the distribution.