Appendix

Appendix A. Discretization of linear stochastic state-space model [72][73]

A linear stochastic state space model (Eq. A.1) can be discretized as (Eq. A.2) assuming zero-order hold. In this Appendix, we will present a discretization of both the state transition and measurement process as a reference. However, in our model, discretized measurement process is directly used.

$$\begin{aligned} \dot{\mathbf{x}}_t &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \boldsymbol{\sigma}_x\dot{\omega} \text{ (state transition process)} \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t + \boldsymbol{\sigma}_y\dot{\omega} \text{ (measurement process)} \end{aligned} \tag{A.1}$$

$$\begin{aligned} \mathbf{x}_{t_k+1} &= \mathbf{A}_\mathrm{d}\mathbf{x}_{t_k} + \mathbf{B}_\mathrm{d}\mathbf{u}_{t_k} + \boldsymbol{\varepsilon}_{x,t_k} \text{ where } \boldsymbol{\varepsilon}_{x,t_k} \sim \boldsymbol{\mathcal{M}\mathcal{V}}\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{d},x}\right) \\ \mathbf{y}_{t_k} &= \mathbf{C}_\mathrm{d}\mathbf{x}_{t_k} + \mathbf{D}_\mathrm{d}\mathbf{u}_{t_k} + \boldsymbol{\varepsilon}_{y,t_k} \text{ where } \boldsymbol{\varepsilon}_{y,t_k} \sim \boldsymbol{\mathcal{M}\mathcal{V}}\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{d},y}\right) \end{aligned} \tag{A.2}$$

where $\mathbf{x} \in \mathbb{R}^{n_x}$ is a vector of states, $\mathbf{y} \in \mathbb{R}^{n_y}$ is a vector of measurements, $\mathbf{A} \in \mathbb{R}^{n_x \times n_x}$ and $\mathbf{B} \in \mathbb{R}^{n_x \times n_u}$ are system parameters in the state transition process, $\mathbf{C} \in \mathbb{R}^{n_y \times n_x}$ and $\mathbf{D} \in \mathbb{R}^{n_y \times n_u}$ are system parameters in the measurement process, t is continuous time, t_k is discrete time, $\dot{\omega}$ is a standard Wiener process, $\boldsymbol{\sigma}_x^2$ and $\boldsymbol{\sigma}_y^2$ are noise variances of state transition and measurement process, $\mathcal{M}\boldsymbol{\mathcal{V}}\mathcal{N}$ is a multivariate normal distribution, subscript d indicates discretization, and $\boldsymbol{\Sigma}_{\mathbf{d},x}$ and $\boldsymbol{\Sigma}_{\mathbf{d},y}$ are noise covariance matrixes of $\mathbf{x}$ and $\mathbf{y}$ after discretization.

$\mathbf{A}_\mathrm{d}$ and $\mathbf{B}_\mathrm{d}$ can be estimated through the property given in Eq. A.3.

$$\exp\begin{bmatrix} \mathbf{A}\Delta t & \mathbf{B}\Delta t \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_\mathrm{d} & \mathbf{B}_\mathrm{d} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{A.3}$$

where exp is matrix exponential, Δt is sampling time for the discrete time.

For the measurement process, $\mathbf{C} = \mathbf{C}_\mathrm{d}$ and $\mathbf{D} = \mathbf{D}_\mathrm{d}$. $\boldsymbol{\Sigma}_{\mathbf{d},x}$ (Eq. A.4 [87]) can be calculated through the property in Eq. (A.5) [88]. Finally, $\boldsymbol{\Sigma}_{\mathbf{d},y}$ is simply $\boldsymbol{\Sigma}_{\mathbf{d},y} = \mathrm{diag}\left[\boldsymbol{\sigma}_y\right]/\Delta \mathrm{t}$ (where diag is diagonal matrix

$$\boldsymbol{\Sigma}_{\mathbf{d},x} = \int_0^{\Delta t} \exp\left(\mathbf{A}(\Delta t - \tau)\right) \mathrm{diag}(\boldsymbol{\sigma}_x) \exp\left(\mathbf{A}^\mathrm{T}(\Delta t - \tau)\right) d\tau \tag{A.4}$$

$$\begin{aligned} &\exp\left(\begin{bmatrix} -\mathbf{A} & \mathrm{diag}(\boldsymbol{\sigma}_x) \\ \mathbf{0} & \mathbf{A}^\mathrm{T} \end{bmatrix} \Delta t\right) = \begin{bmatrix} \cdots & \mathbf{A}_\mathrm{d}^{-1}\boldsymbol{\Sigma}_{\mathbf{d},x} \\ \mathbf{0} & \mathbf{A}_\mathrm{d}^\mathrm{T} \end{bmatrix} \\ &\boldsymbol{\Sigma}_{\mathbf{d},x} = \mathbf{A}_\mathrm{d}\left(\mathbf{A}_\mathrm{d}^{-1}\boldsymbol{\Sigma}_{\mathbf{d},x}\right) \end{aligned} \tag{A.5}$$

Appendix B. Zero-process noise model

In general, three approaches are used to learn parameters in a linear state-space model. First approach is Bayesian approach. With the probabilistic formulation, state-marginalized likelihood (obtained via Kalman filter Eq. B.7) is used to sample posterior distribution of parameters via MCMC. The second approach is to use an optimizer to fine a set of best parameters that maximize the state-marginalized likelihood (hereinafter referred to as a "MLE approach"). These two methods handle an identical problem but solve in a different perspective. The more non-informative prior or the more data are used, their results are getting similar except for MLE is point estimate. The final approach is to find a set of parameters that gives minimum prediction error through optimization (Eq. 14) (hereinafter referred to as a "deterministic approach"). Since the first and second approaches handle a same probabilistic problem, we will show how the second (MLE) and the third approach (deterministic approach) differ in this section. To sum, they give same results when the process (state transition) noises are zero with scaled identity matrix for a measurement noise.

A discretized linear state-space model in Eq. A.2 can be expressed in a probabilistic format as Eq. B.1 (see Appendix A for nomenclature).

$$\begin{aligned}
&P\left(\mathbf{x}_{t_k+1}|\mathbf{x}_{t_k}\right) = \mathcal{MVN}\left(\mathbf{x}_{t_k+1}|f_{\mathrm{d}}\left(\mathbf{x}_{t_k}, \mathbf{u}_{t_k}, \boldsymbol{\theta}\right), \boldsymbol{\Sigma}_{\mathrm{d},x}\right) \\
&P\left(\mathbf{y}_{t_k}|\mathbf{x}_{t_k}\right) = \mathcal{MVN}\left(\mathbf{y}_{t_k}|g_{\mathrm{d}}\left(\mathbf{x}_{t_k}, \mathbf{u}_{t_k}, \boldsymbol{\theta}\right), \boldsymbol{\Sigma}_{\mathrm{d},y}\right) \\
&f_{\mathrm{d}}\left(\mathbf{x}_{t_k}, \mathbf{u}_{t_k}, \boldsymbol{\theta}\right) = \mathbf{A}_{\mathrm{d}}\mathbf{x}_{t_k} + \mathbf{B}_{\mathrm{d}}\mathbf{u}_{t_k} \\
&g_{\mathrm{d}}\left(\mathbf{x}_{t_k}, \mathbf{u}_{t_k}, \boldsymbol{\theta}\right) = \mathbf{C}_{d}\mathbf{x}_{t_k} + \mathbf{D}_{\mathrm{d}}\mathbf{u}_{t_k}
\end{aligned} \tag{B.1}$$

The joint likelihood of hidden states and parameter can be written as Eq. B.2 due to its Markov structure.

$$\begin{aligned}
P\left(\mathbf{x}_{1:t_k}, \boldsymbol{\theta}|\mathbf{y}_{1:t_k}\right) &\propto P\left(\mathbf{y}_{t_k}|\mathbf{x}_{1:t_k}, \boldsymbol{\theta}, \mathbf{y}_{1:t_k-1}\right)P\left(\mathbf{x}_{1:t_k}, \boldsymbol{\theta}|\mathbf{y}_{1:t_k-1}\right) \\
&\propto P\left(\mathbf{y}_{t_k}|\mathbf{x}_{t_k}, \boldsymbol{\theta}\right)P\left(\mathbf{x}_{t_k}|\mathbf{x}_{t_k-1}, \boldsymbol{\theta}\right)P\left(\mathbf{x}_{1:t_k-1}, \boldsymbol{\theta}|\mathbf{y}_{1:t_k-1}\right) \\
&\propto P(\boldsymbol{\theta})P(\mathbf{x}_1)\prod_{i=2}^{t_k} P(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})P(\mathbf{x}_i|\mathbf{x}_{i-1}, \boldsymbol{\theta})
\end{aligned} \tag{B.2}$$

Since the space of hidden states ($\mathbf{x}_{1:t_k}$) is proportional to the time ($1{:}\, t_k$), it is convenient to marginalize hidden space to learn parameter. Before deriving the state-marginalized likelihood, predicted and filtered states are expressed as Eq. B.3.

$$\begin{aligned}
&P\left(\mathbf{x}_{t_k}|\mathbf{y}_{1:t_k-1}\right) = \mathcal{MVN}\left(\mathbf{x}_{t_k+1}|\boldsymbol{\mu}_{x_{\mathrm{p}},t_k}, \mathbf{P}_{\mathrm{p},t_k}\right) \text{ (predicted states)} \\
&P\left(\mathbf{x}_{t_k}|\mathbf{y}_{1:t_k}\right) = \mathcal{MVN}\left(\mathbf{x}_{t_k}|\boldsymbol{\mu}_{x_{\mathrm{f}},t_k}, \mathbf{P}_{\mathrm{f},t_k}\right) \text{ (filtered states)}
\end{aligned} \tag{B.3}$$

The states-marginalized likelihood ($P\left(\mathbf{y}_{1:t_k}|\boldsymbol{\theta}\right)$) can be expressed as Eq. B.4[89][90].

$$\begin{aligned}
P\left(\mathbf{y}_{1:t_k}|\boldsymbol{\theta}\right) &= \int P(\mathbf{x}_1, \mathbf{y}_1|\boldsymbol{\theta})\, d\mathbf{x}_1 \prod_{i=2}^{t_k} \int P(\mathbf{x}_i, \mathbf{y}_i|\boldsymbol{\theta}, \mathbf{y}_1, \dots, \mathbf{y}_{i-1})\, d\mathbf{x}_i \\
&= \int P(\mathbf{y}_1|\mathbf{x}_1, \boldsymbol{\theta})P(\mathbf{x}_1|\boldsymbol{\theta})\, d\mathbf{x}_1 \prod_{i=2}^{t_k} \int P(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})P(\mathbf{x}_i|\boldsymbol{\theta}, \mathbf{y}_1, \dots, \mathbf{y}_{i-1})\, d\mathbf{x}_i \\
&= \prod_{i=1}^{t_k} \int \mathcal{MVN}\left(\mathbf{y}_i|\mathbf{C}_{\mathrm{d}}\boldsymbol{\mu}_{x_f,i} + \mathbf{D}_{\mathrm{d}}\mathbf{u}_i, \boldsymbol{\Sigma}_{\mathrm{d},y}\right)\mathcal{MVN}\left(\mathbf{x}_i|\boldsymbol{\mu}_{\mathbf{x}_{\mathrm{p}},i}, \mathbf{P}_{\mathrm{p},i}\right) d\mathbf{x}_i = \prod_{i=1}^{t_k} \mathcal{MVN}\left(\mathbf{y}_i|\mathbf{C}_{\mathrm{d}}\boldsymbol{\mu}_{x_{\mathrm{p}},i}, \mathbf{S}_i\right)
\end{aligned} \tag{B.4}$$

where $\mathbf{x}_1 \sim \mathcal{MVN}\left(\boldsymbol{\mu}_{x_1}, \mathbf{P}_1\right)$, i is time. $\mathbf{P}_{\mathrm{p},i}$ and $\mathbf{P}_{\mathrm{f},i}$ are predicted and filtered state covariance, respectively. $\boldsymbol{\mu}_{x_{\mathrm{p}},i}$ and $\boldsymbol{\mu}_{x_{\mathrm{f}},i}$ are predicted and filtered state mean, respectively. $\mathbf{S}_i$ is innovation covariance. This can be estimated through the Kalman recursion (Eq. B.5).

$$\begin{aligned}
&\text{Predicted state mean: } \boldsymbol{\mu}_{x_{\mathrm{p}},t_k} = \mathbf{A}_{\mathrm{d}}\boldsymbol{\mu}_{x_{\mathrm{f}},t_k-1} + \mathbf{B}_{\mathrm{d}}\mathbf{u}_{t_k-1}\\
&\text{Predicted state covariance: } \mathbf{P}_{\mathrm{p},t_k} = \mathbf{A}_{\mathrm{d}}\mathbf{P}_{\mathrm{f},t_k-1}\mathbf{A}_{\mathrm{d}}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\mathrm{d},x}\\
&\text{Innovation: } \boldsymbol{\varepsilon}_{t_k} = \mathbf{y}_{t_k} - \mathbf{C}_{\mathrm{d}}\,\boldsymbol{\mu}_{x_{\mathrm{p}},t_k}\\
&\text{Innovation covariance: } \mathbf{S}_{t_k} = \mathbf{C}_{\mathrm{d}}\mathbf{P}_{\mathrm{p},t_k-1}\mathbf{C}_{\mathrm{d}}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\mathrm{d},y}\\
&\text{Kalman gain: } \mathbf{K}_{t_k} = \mathbf{P}_{\mathrm{p},t_k}\mathbf{C}_{\mathrm{d}}^{\mathrm{T}}\mathbf{S}_{t_k}^{-1}\\
&\text{Filtered states mean: } \boldsymbol{\mu}_{x_{\mathrm{f}},i} = \boldsymbol{\mu}_{x_{\mathrm{p}},i} + \mathbf{K}_{t_k}\boldsymbol{\varepsilon}_{t_k}\\
&\text{Filtered states cavariance: } \mathbf{P}_{\mathrm{f},t_k} = (\mathbf{I} - \mathbf{K}_{t_k}\mathbf{C}_{\mathrm{d}})\mathbf{P}_{\mathrm{p},t_k}
\end{aligned} \qquad \text{(B.5)}$$

Therefore, the state-marginalized likelihood shows a property in Eq. B.6.

$$P\left(\mathbf{y}_{1:t_k}|\boldsymbol{\theta}\right) = \prod_{i=1}^{t_k} \mathcal{MVN}\left(\mathbf{y}_i|\mathbf{C}_{\mathrm{d}}\boldsymbol{\mu}_{\mathbf{x}_{\mathrm{p}},i}, \mathbf{S}_i\right) = \prod_{i=1}^{t_k} \frac{\exp\left(-\frac{1}{2}\left(\mathbf{y}_i - \mathbf{C}_{\mathrm{d}}\,\boldsymbol{\mu}_{x_{\mathrm{p}},i}\right)^{\mathrm{T}}\mathbf{S}_i^{-1}\left(\mathbf{y}_i - \mathbf{C}_{\mathrm{d}}\,\boldsymbol{\mu}_{x_{\mathrm{p}},i}\right)\right)}{\sqrt{2\pi^{n_y}|\mathbf{S}_i|}} \qquad \text{(B.6)}$$

The log-likelihood of Eq. B.6 is

$$\log P\left(\mathbf{y}_{1:t_k}|\boldsymbol{\theta}\right) = \sum_{i=1}^{t_k}\left[-\frac{1}{2}\left(\boldsymbol{\varepsilon}_i^{\mathrm{T}}\mathbf{S}_i^{-1}\boldsymbol{\varepsilon}_i + \log|\mathbf{S}_i| + n_y\log 2\pi\right)\right] \qquad \text{(B.7)}$$

In MLE approach, an optimizer will find best set of parameters that maximize this log-likelihood. This can be viewed as one-step ahead prediction method because the log-likelihood of each timestep is calculated one-step ahead prediction from the filtered states of the previous time. To make a comparison with deterministic approach, let's define a zero-process noise case (i.e., $\mathbf{P}_1 = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\mathrm{d,x}} = \mathbf{0}$). From the Kalman recursion in Eq. B.5, $\mathbf{S}_{t_k} = \boldsymbol{\Sigma}_{\mathrm{d,y}}$ and $\boldsymbol{\mu}_{\mathbf{x}_{\mathrm{f}},t_k} = \boldsymbol{\mu}_{\mathbf{x}_{\mathrm{p}},t_k} = \mathbf{A}_{\mathrm{d}}\mathbf{x}_{t_k-1} + \mathbf{B}_{\mathrm{d}}\mathbf{u}_{t_k-1}$. Then, its log-likelihood is Eq. B.8.

$$\log P\left(\mathbf{y}_{1:t_k}|\boldsymbol{\theta}\right) = \sum_{i=1}^{t_k}\left[-\frac{1}{2}\left(\boldsymbol{\varepsilon}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{d},y}^{-1}\boldsymbol{\varepsilon}_i + \log\left|\boldsymbol{\Sigma}_{\mathrm{d},y}\right| + n_y\log 2\pi\right)\right] \qquad \text{(B.8)}$$

Because $\boldsymbol{\varepsilon}_{t_k}$ is a prediction error of n-step ahead prediction ($\mathbf{y}_{t_k} - \mathbf{C}_{\mathrm{d}}\,\boldsymbol{\mu}_{\mathbf{x}_{\mathrm{p}},t_k}$), when $\boldsymbol{\Sigma}_{\mathrm{d},y}$ is a scaled identity matrix (i.e., constant × $\boldsymbol{I}_{n_y}$), the $\boldsymbol{\varepsilon}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{d},y}^{-1}\boldsymbol{\varepsilon}_i$ part in Eq. B.8 is square of prediction error in optimization approach with l_2 norm as a cost function (Eq. 14). Therefore, the solution of MLE approach is same with deterministic approach when zero-process noise with scaled identity matrix for a measurement noise.