PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 13: SEQUENTIAL DATA

Learning Objectives

1、What are hidden Markov models (HMMs)?

2、What is the EM scheme for HMMs?

3、What are Forward-Backward and Sum-Product Algorithms?

4、What are Viterbi and Max-Product Algorithms?

5、What are linear dynamic systems?

6、What are Kalman and particle filters?

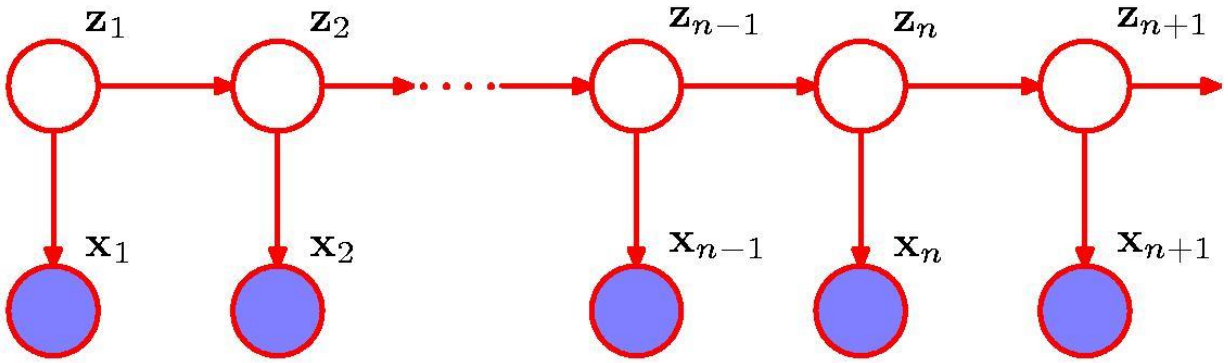
7、How to learn linear dynamic system models?

8、What are RNN and LSTM?

Outlines

- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi and Max-Product Algorithms
- Linear Dynamics Systems
- Kalman and Particle Filters
- RNN and LSTM

Stochastic Linear Dynamical Systems



$$p(\mathbf{z}_n|\mathbf{z}_{n-1}) = \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \mathbf{\Gamma})$$

$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \mathbf{\Sigma})$$

$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0)$$

$$\begin{aligned}
\mathbf{z}_n &= \mathbf{A}\mathbf{z}_{n-1} + \mathbf{w}_n & \mathbf{w} &\sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{\Gamma}) \\
\mathbf{x}_n &= \mathbf{C}\mathbf{z}_n + \mathbf{v}_n & \mathbf{v} &\sim \mathcal{N}(\mathbf{v}|\mathbf{0}, \mathbf{\Sigma}) \\
\mathbf{z}_1 &= \boldsymbol{\mu}_0 + \mathbf{u} & \mathbf{u} &\sim \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{V}_0)
\end{aligned}$$

Inference Problem

- Finding the marginal distributions for the latent variables conditional on the observation sequence.

$$\widehat{\alpha}(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n)$$

$$\widehat{\beta}(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N)$$

$$\gamma(\mathbf{z}_n) = \widehat{\alpha}(\mathbf{z}_n)\widehat{\beta}(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n)$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = (c_n)^{-1}\widehat{\alpha}(\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{-1})\widehat{\beta}(\mathbf{z}_n)$$

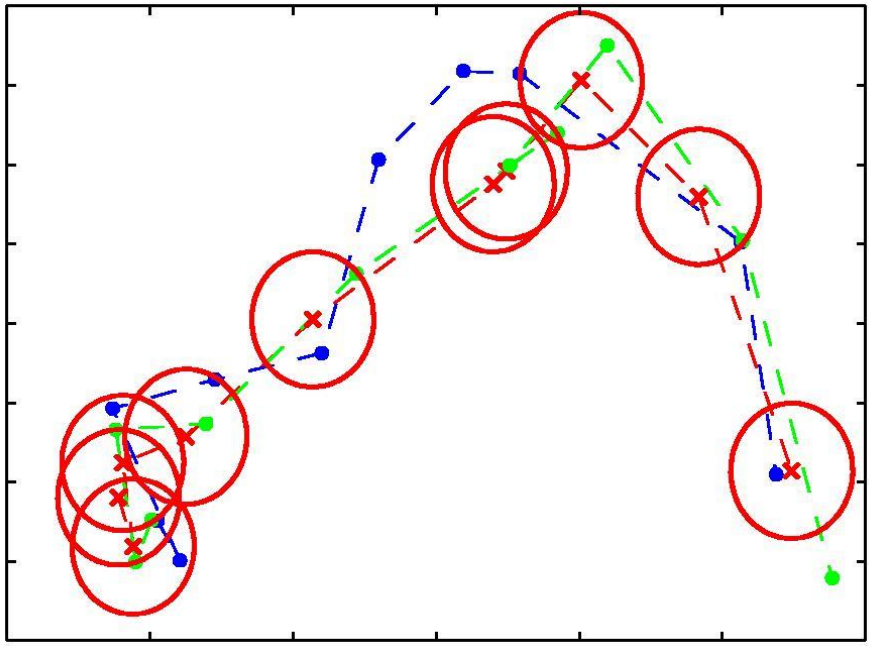
$$c_n = p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$$

Outlines

- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi and Max-Product Algorithms
- Linear Dynamics Systems
- Kalman Filters and LDS Learning
- RNN and LSTM

Application: Tracking an Moving Object

- One of the most important application of the Kalman filter.



An illustration of a linear dynamical system used to track a moving object.

Blue: $\mathbf{Z}_n$

Green: $\mathbf{X}_n$

Red: $\mathbf{Z}_n \left| \mathbf{X}_1, \ldots, \mathbf{X}_n \right.$

Mean and Variance of Kalman Filter

- Kalman filter equations

$$
\begin{aligned}
\boldsymbol{\mu}_n &= \mathbf{A}\boldsymbol{\mu}_{n-1} + \mathbf{K}_n(\mathbf{x}_n - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}) \\
\mathbf{V}_n &= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_{n-1} \\
c_n &= \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma})
\end{aligned}
$$

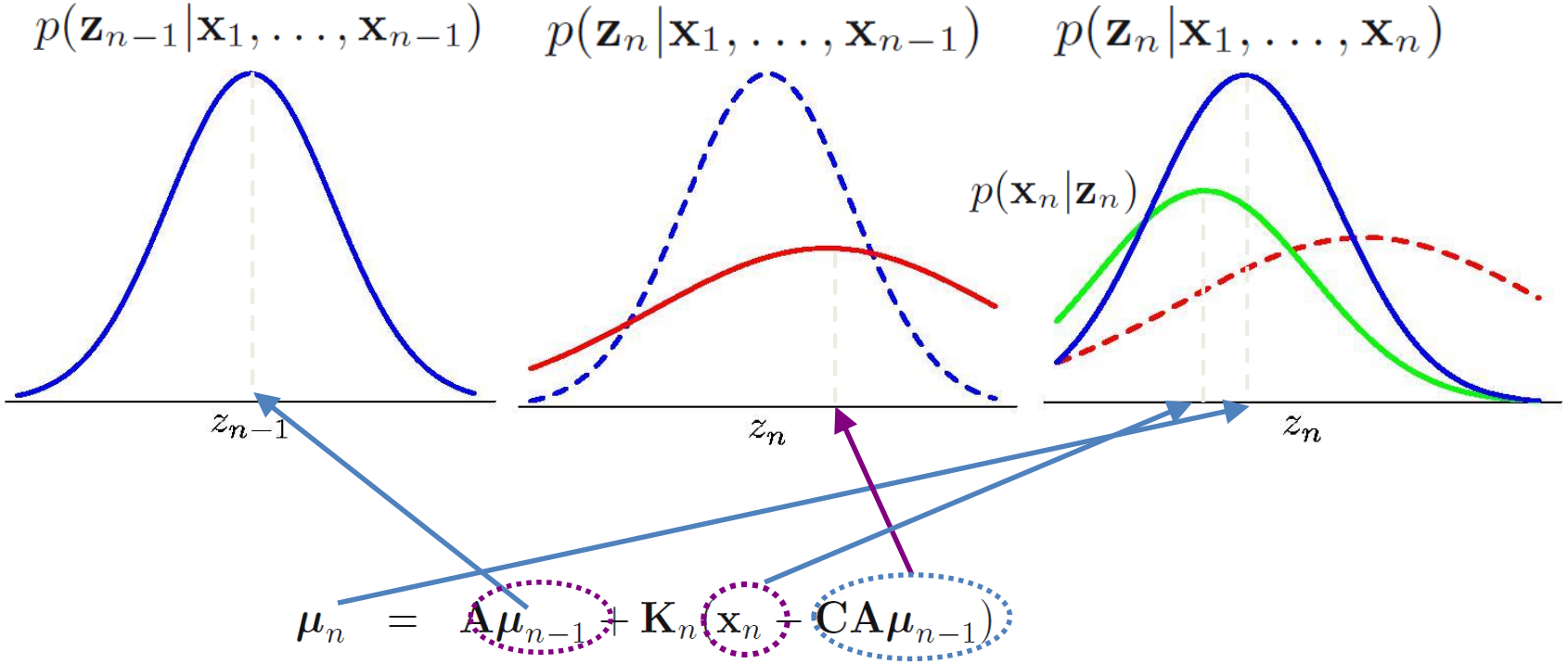
$$
\mathbf{P}_{n-1} = \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Gamma}
$$

$$
\mathbf{K}_n = \mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}\right)^{-1}
$$

$$
\begin{aligned}
\widehat{\alpha}(\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) \\
c_n &= p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})
\end{aligned}
$$

Interpretation of Kalman Filters

- Kalman filter as a process of
  - ✓ Making successive *predictions*
  - ✓ *Correcting* the predictions using new observations.



Mean and Variance of Full Kalman Filter

- Full Kalman filter equations

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_n &= \boldsymbol{\mu}_n + \mathbf{J}_n\left(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{A}\boldsymbol{\mu}_N\right) \\ \widehat{\mathbf{V}}_n &= \mathbf{V}_n + \mathbf{J}_n\left(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\right)\mathbf{J}_n^{\mathrm{T}}\end{aligned}$$

$$\mathbf{J}_n = \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\left(\mathbf{P}_n\right)^{-1}$$

$$\mathbf{A}\mathbf{V}_n = \mathbf{P}_n\mathbf{J}_n^{\mathrm{T}}$$

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_1, \ldots, \mathbf{x}_N) = \mathcal{N}(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n)$$

Covariance of Sequential States

- Joint posterior of sequential states is Gaussian

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{z}_{n-1}, \mathbf{z}_n)p(\mathbf{z}_{n-1}, \mathbf{z}_n)}{p(\mathbf{X})}$$

$$= \frac{\mathcal{N}(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \mathbf{V}_{n-1})\mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma})\mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n)}{c_n\widehat{\alpha}(\mathbf{z}_n)}$$

$$\boxed{\mathrm{cov}[\mathbf{z}_n, \mathbf{z}_{n-1}] = \mathbf{J}_{n-1}\widehat{\mathbf{V}}_n}$$

Learning Problem

- Determining the parameters $\vartheta = \{\mathbf{A}, \Gamma, \mathbf{C}, \Sigma, \mu_0, \mathbf{V}_0\}$ using the *EM algorithm*.

- The complete data $\{\mathbf{X}, \mathbf{Z}\}$ log likelihood function

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=2}^{N} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}, \boldsymbol{\Gamma}) + \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{C}, \boldsymbol{\Sigma})$$

Expectation of Latent Variables

- The expectation of latent variables

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{z}_n\right] &= \widehat{\boldsymbol{\mu}}_n \\
\mathbb{E}\left[\mathbf{z}_n \mathbf{z}_{n-1}^{\mathrm{T}}\right] &= \mathbf{J}_{n-1}\widehat{\mathbf{V}}_n + \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_{n-1}^{\mathrm{T}} \\
\mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^{\mathrm{T}}\right] &= \widehat{\mathbf{V}}_n + \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_n^{\mathrm{T}}
\end{aligned}
$$

$$
\operatorname{cov}\left[\mathbf{z}_n, \mathbf{z}_{n-1}\right] = \mathbf{J}_{n-1}\widehat{\mathbf{V}}_n
$$

Expectation of Log Likelihood Function

- The expectation of the log likelihood function with respect to $p(\mathrm{Z} \mid \mathrm{X}, \theta^{\mathrm{old}})$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{old}}} \left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right]$$

$$= -\frac{1}{2} \ln |\mathbf{V}_0| - \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{old}}} \left[\frac{1}{2} (\mathbf{z}_1 - \boldsymbol{\mu}_0)^{\mathrm{T}} \mathbf{V}_0^{-1} (\mathbf{z}_1 - \boldsymbol{\mu}_0) \right] + \mathrm{const}$$

$$= -\frac{N-1}{2} \ln |\boldsymbol{\Gamma}| - \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{old}}} \left[\frac{1}{2} \sum_{n=2}^{N} (\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} (\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1}) \right] + \mathrm{const}$$

$$= -\frac{N}{2} \ln |\boldsymbol{\Sigma}| - \mathbb{E}_{\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{old}}} \left[\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n) \right] + \mathrm{const}.$$

Maximization of LSD Parameters

$$\boldsymbol{\mu}_0^{\text{new}} = \mathbb{E}[\mathbf{z}_1]$$

$$\mathbf{V}_0^{\text{new}} = \mathbb{E}[\mathbf{z}_1\mathbf{z}_1^{\mathrm{T}}] - \mathbb{E}[\mathbf{z}_1]\mathbb{E}[\mathbf{z}_1^{\mathrm{T}}]$$

$$\mathbf{A}^{\text{new}} = \left(\sum_{n=2}^{N}\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right]\right)\left(\sum_{n=2}^{N}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right]\right)^{-1}$$

$$\boldsymbol{\Gamma}^{\text{new}} = \frac{1}{N-1}\sum_{n=2}^{N}\left\{\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right] - \mathbf{A}^{\text{new}}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_n^{\mathrm{T}}\right] - \mathbb{E}\left[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right]\mathbf{A}^{\text{new}} + \mathbf{A}^{\text{new}}\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right](\mathbf{A}^{\text{new}})^{\mathrm{T}}\right\}$$

$$\mathbf{C}^{\text{new}} = \left(\sum_{n=1}^{N}\mathbf{x}_n\mathbb{E}\left[\mathbf{z}_n^{\mathrm{T}}\right]\right)\left(\sum_{n=1}^{N}\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\right)^{-1}$$

$$\boldsymbol{\Sigma}^{\text{new}} = \frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \mathbf{C}^{\text{new}}\mathbb{E}\left[\mathbf{z}_n\right]\mathbf{x}_n^{\mathrm{T}} - \mathbf{x}_n\mathbb{E}\left[\mathbf{z}_n^{\mathrm{T}}\right]\mathbf{C}^{\text{new}} + \mathbf{C}^{\text{new}}\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right]\mathbf{C}^{\text{new}}\right\}$$

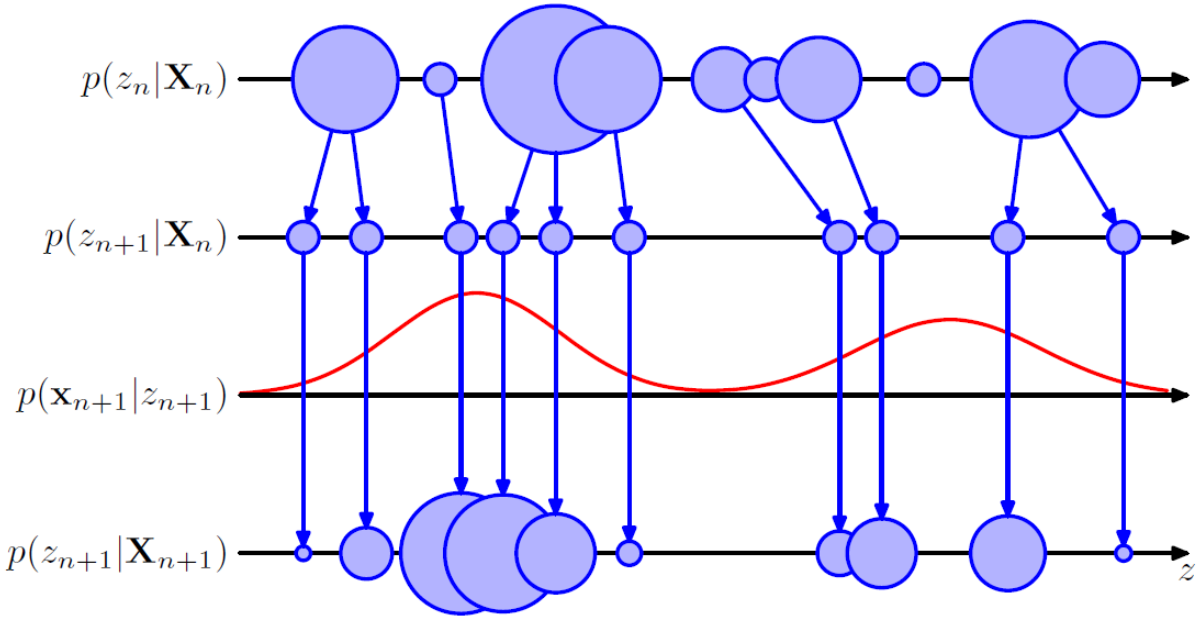
Extensions of LDS

- Problem: Beyond the linear-Gaussian assumption.
  - ✓ Considerable interest in *extending the basic linear dynamical system* in order to increase its capabilities.
  - ✓ Gaussian $p(\mathbf{z}_n | \mathbf{x}_n)$ – A significant limitation.

- Some extensions
  - ✓ Gaussian mixture $p(\mathbf{z}_n)$.
  - ✓ The extended Kalman filter.
  - ✓ The switching state space model
  - ✓ The switching hidden Markov model

Particle filters

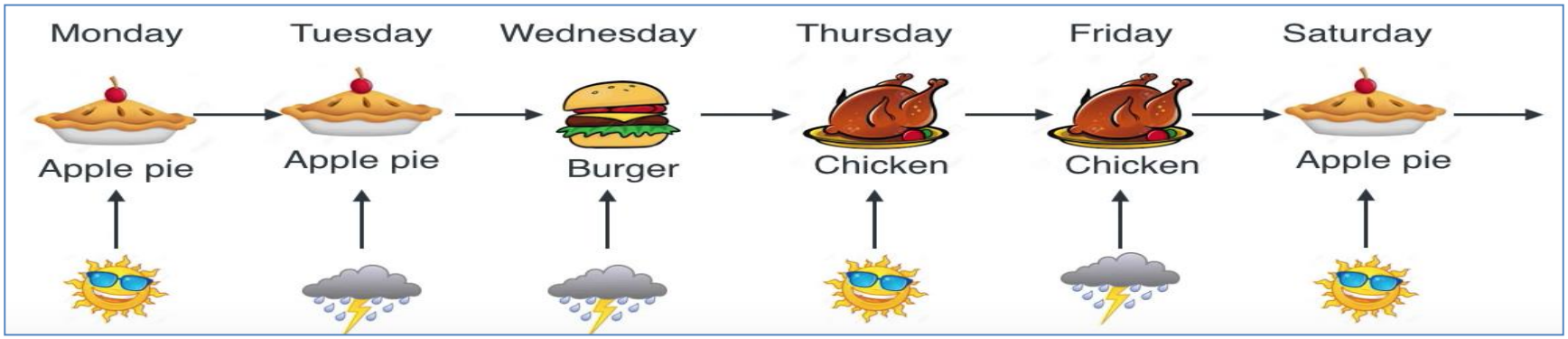
- Non-Gaussian emission density $p(\mathbf{x}_n|\mathbf{z}_n)$
  - ✓ non-Gaussian $p(z_n \mid x_1, \ldots, x_n)$
  - ✓ mathematically intractable integral
- Sampling-importance-resampling

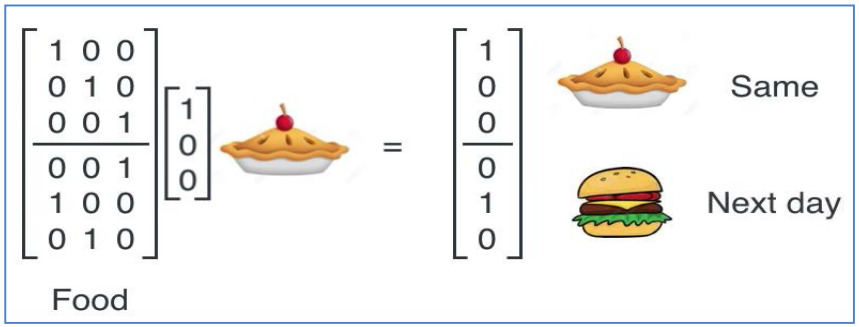


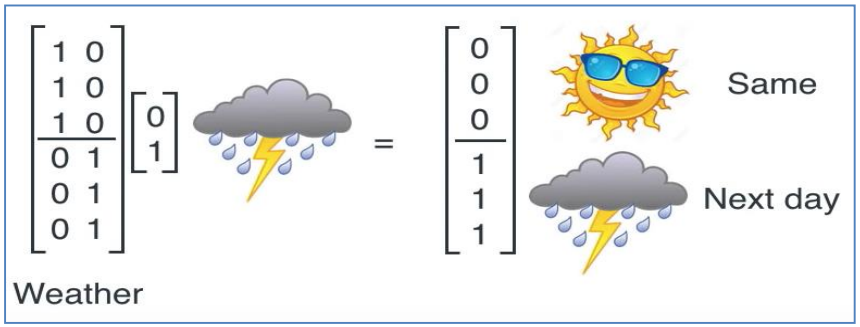
Outlines

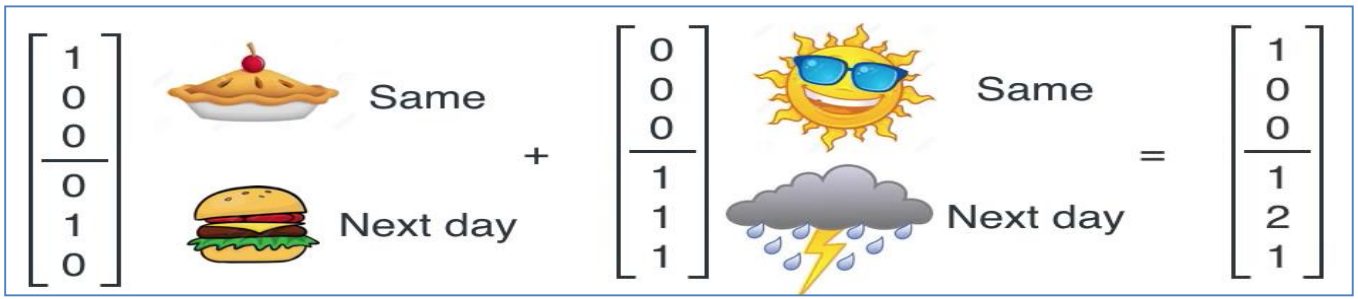
- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi and Max-Product Algorithms
- Linear Dynamics Systems
- Kalman Filters and LDS Learning
- RNN and LSTM

Complicated Sequential Data

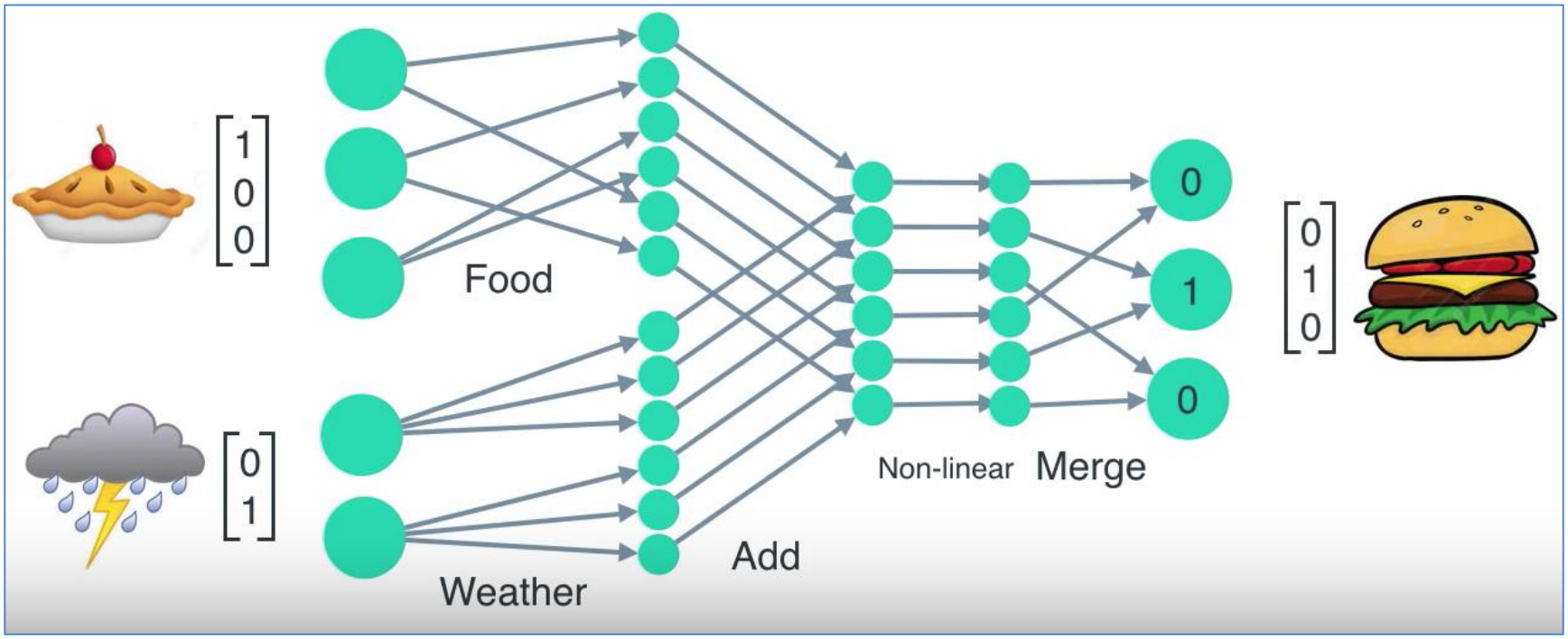




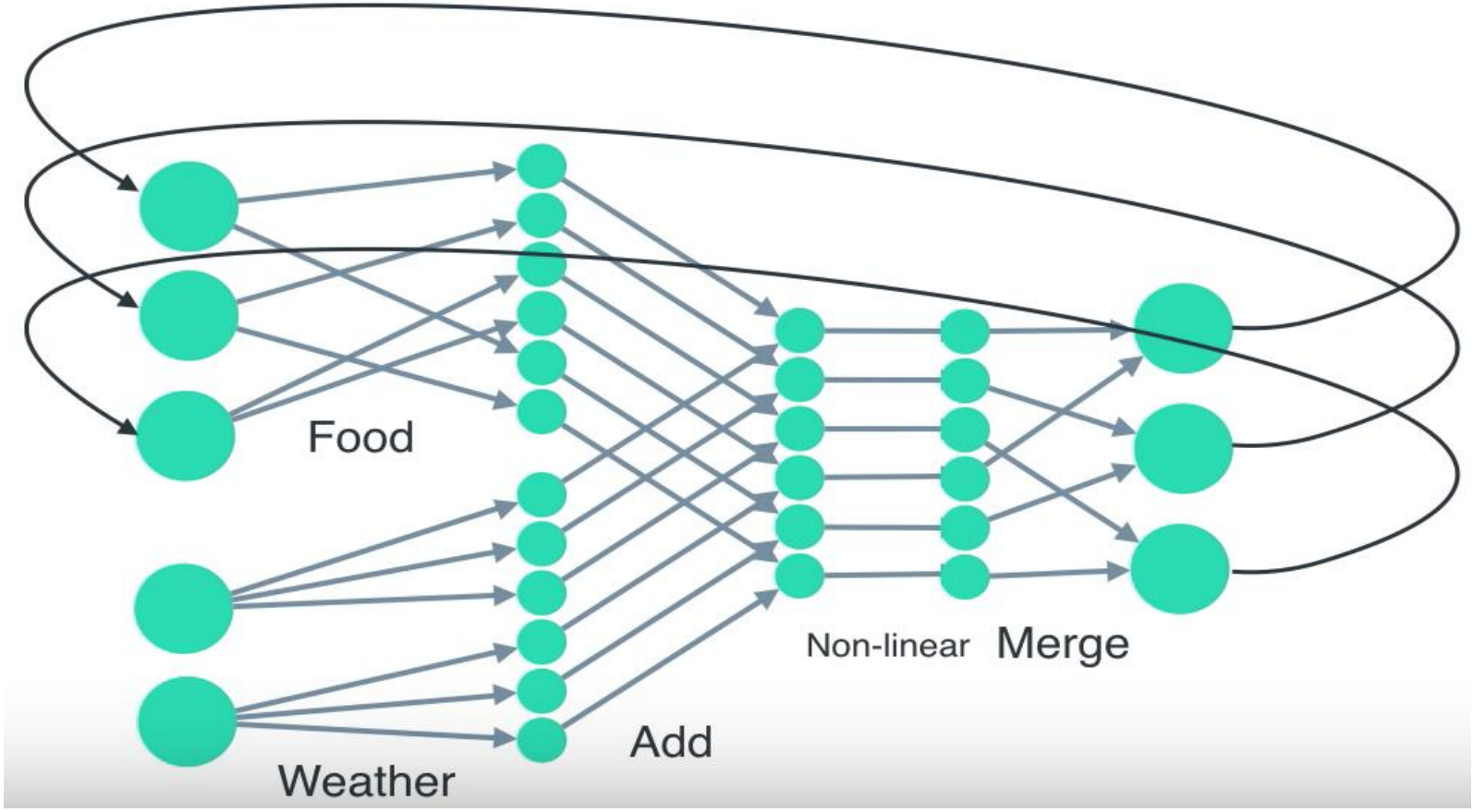




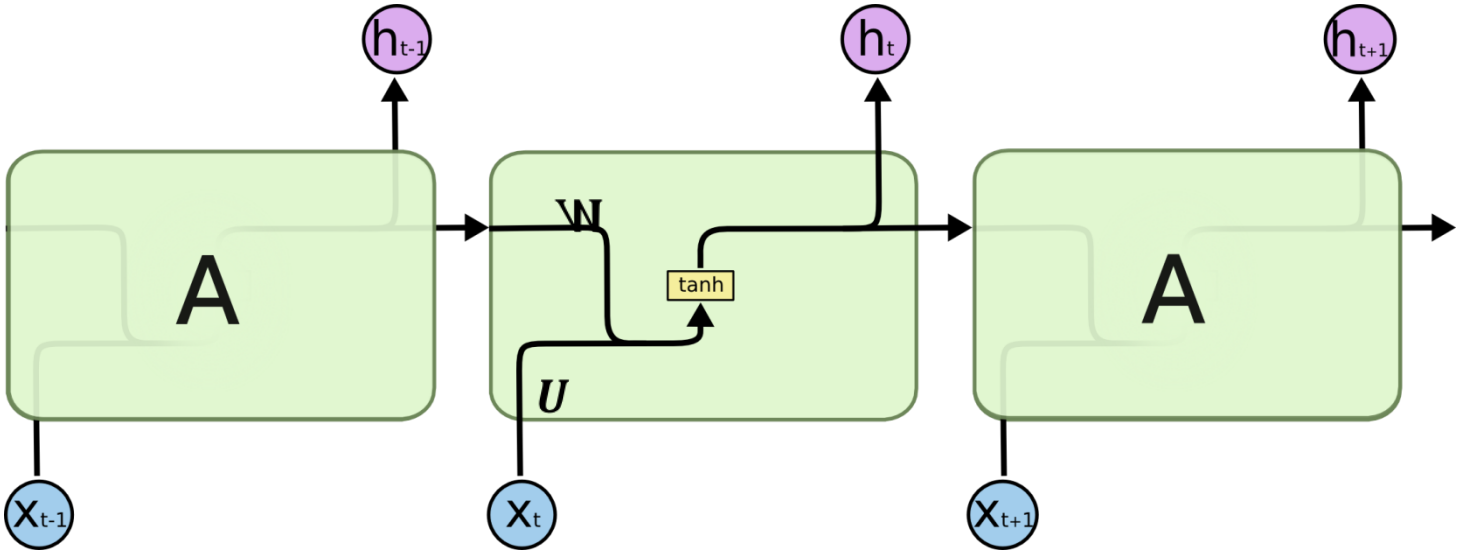
Neural Network



Recurrent Neural Network



Standard RNN Modules

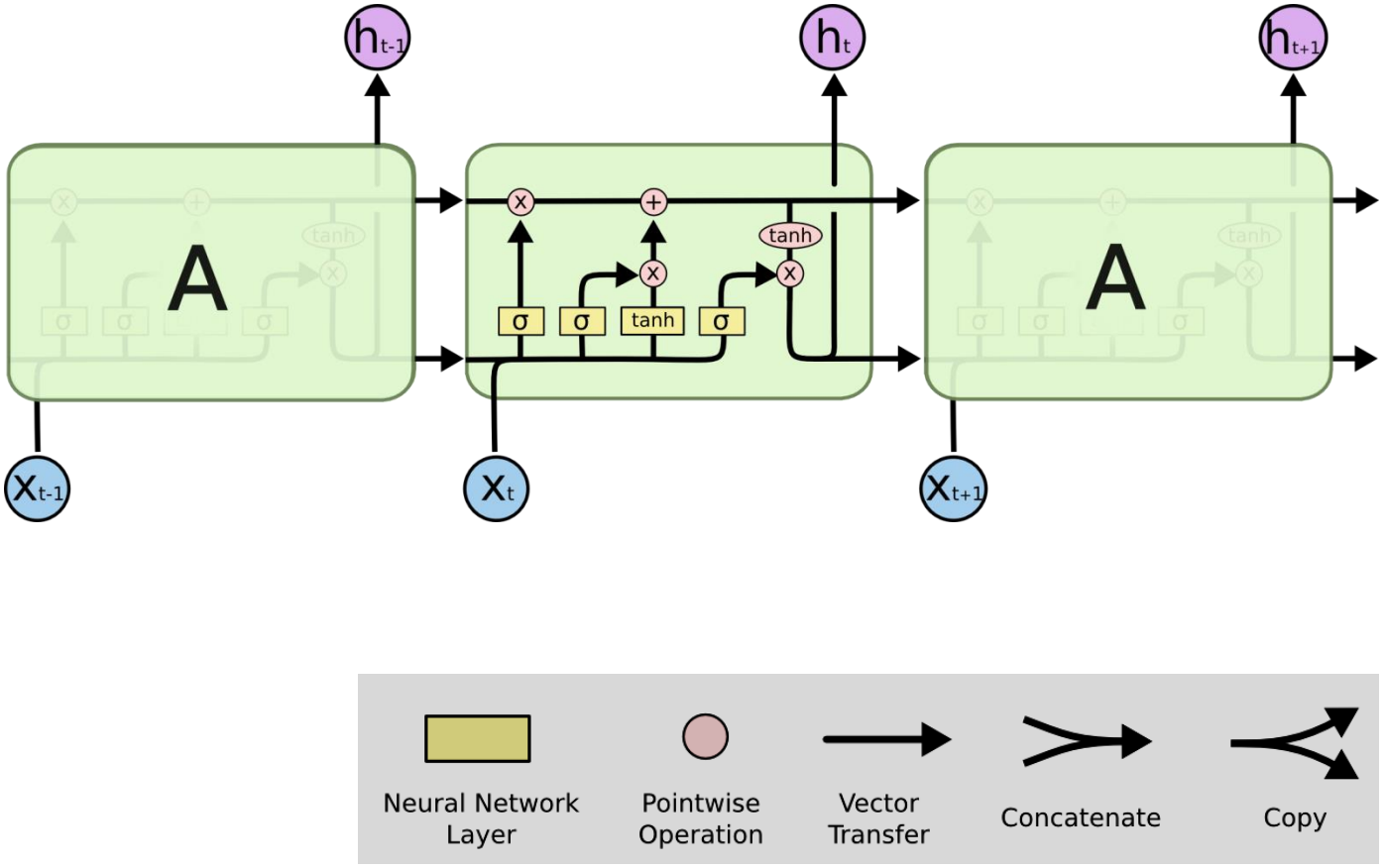


$$\boldsymbol{h_t = f(Ux_t + Wh_{t-1} + b)}$$

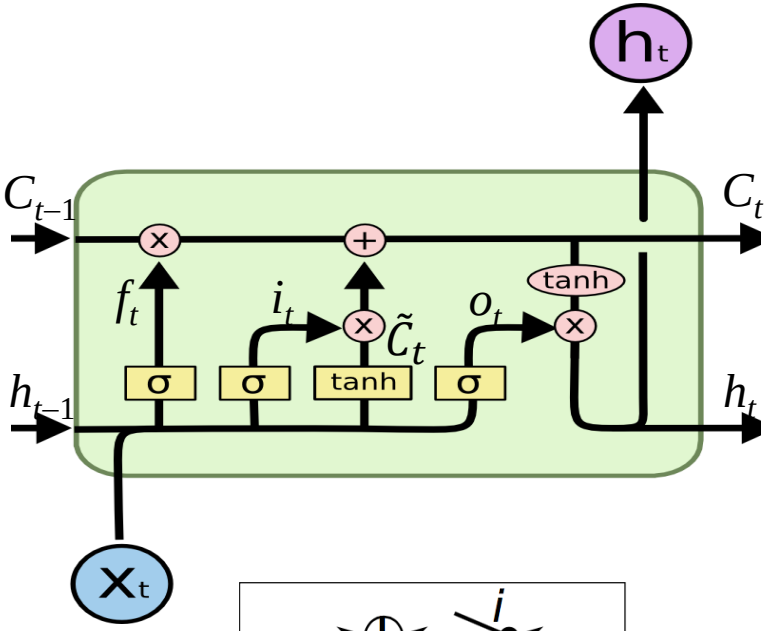
$\mathbf{h}_t$: output

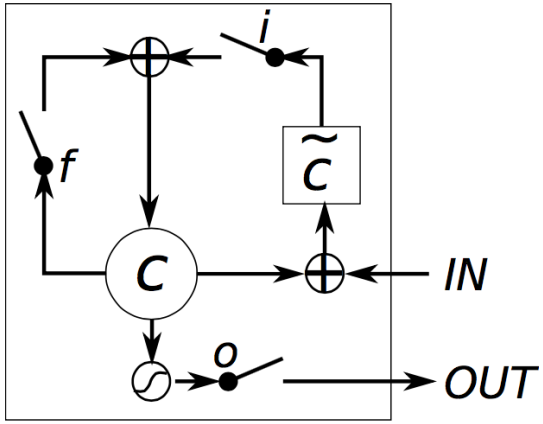
$\boldsymbol{x}_t$: input

Long Short Term Memory



Long Short Term Memory





$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

C_t: cell state
$\tilde{C}_t$: cell state prediction
f_t : forget gate
i_t : input gate
o_t : output gate
h_t : output
x_t : input

Summary

- Hidden Markov Models
- Maximum Likelihood and EM for HMM
- Forward-Backward and Sum-Product Algorithms
- Viterbi and Max-Product Algorithms
- Linear Dynamics Systems
- Kalman Filters and LDS Learning
- RNN and LSTM