15: Probabilistic Reasoning Over Time

Modeling Uncertainty over Time

- Setting
  - X_t - a set of unobserved state variables at time *t*.
  - E_t - a set of observable evidence variables for time *t*.
  - *a:b* – denotes an interval from *a* to *b*.
- **Stationary Process –** process of change that is governed by laws that do not change over time.
- **Markov Assumption –** current state depends only on a *finite* history of previous states. Processes satisfying this assumption are *Markov Processes (Chains)*.
  - **transition model** – law describing how state changes over time.
    $$P\left(X_t \mid X_{0:t-1}\right) = P\left(X_t \mid X_\alpha\right) \text{ where } \alpha \subseteq \{1 \ldots t-1\}$$
  - **first-order Markov Process** – current state is solely dependent on the previous state
    - transition model: $P\left(X_t \mid X_{t-1}\right)$
- We assume the evidence variables at time *t* depend only on the current state.
  - **sensor model** – law describing how the evidence depends on the state.
    $$P\left(E_t \mid X_{0:t}, E_{0:t-1}\right) = P\left(E_t \mid X_t\right)$$
- prior probability for the initial state: $P\left(X_0\right)$
- complete joint
  $$P\left(X_{0:T}, E_{1:T}\right) = P\left(X_0\right)\prod_{t=1}^{T} P\left(X_t \mid X_{t-1}\right) P\left(E_t \mid X_t\right)$$
- Ways to deal with inaccurate Markov modeling:
  1. Increase the order of the Markov process
  2. Increase the set of state variables

Filter (monitoring) – the task of computing the *belief state* – the posterior distribution of the current state given all evidence; $P\left(X_T \mid e_{1:T}\right)$.

- Recursive estimation – forward chaining.
  $$P\left(X_t \mid e_{1:t}\right) \propto P\left(e_t \mid X_t\right) \sum_{X_{t-1}} P\left(X_t \mid X_{t-1}\right) \underbrace{P\left(X_{t-1} \mid e_{1:t-1}\right)}_{\text{recursive estimate}}$$
  $$f_{1:t} \propto FORWARD\left(f_{1:t-1}, e_t\right)$$
- When the state variables are discrete, this update is constant in space and time.
- *Likelihood* $P\left(e_{1:T}\right)$ can be calculated by a likelihood message: $l_{1:t} = P\left(X_t, e_{1:t}\right)$:
  $$L_{1:T} = \sum_{X_T} l_{1:T}\left(X_T, e_{1:T}\right)$$

Prediction – task of computing the posterior distribution over a *future* state, given all evidence; $P\left(X_{T+k} \mid e_{1:T}\right)$ where *k > 0*.

- This is equivalent to filtering without new evidence. Hence, we can easily derive the following update:

$$P\left(X_{T+k} \mid e_{1:T}\right) = \sum_{X_{t+k}} P\left(X_{T+k} \mid X_{T+k-1}\right) \underbrace{P\left(X_{T+k-1} \mid e_{1:T}\right)}_{\text{recursive estimate}}$$

- **stationary distribution –** The fixed point of the Markov process that is approached upon successive applications of the transition model.
    - **mixing time –** the amount of time required to reach stationarity.
    - Prediction is doomed to failure for future times more than a small fraction of the mixing time.

Smoothing (hindsight) – task of computing posterior distribution for a *past* state, given all evidence; $P\left(X_k \mid e_{1:T}\right)$ where $0 \le k < T$.

- Accounting for hindsight is done with an additional backwards message:

$$P\left(X_k \mid e_{1:T}\right) \propto \underbrace{P\left(X_k \mid e_{1:k}\right)}_{f_{1:k}} \underbrace{P\left(e_{k+1:T} \mid X_k\right)}_{b_{k+1:T}}$$

$$b_{k+1:T} = \sum_{X_{k+1}} P\left(e_{k+1} \mid X_{k+1}\right) P\left(X_{k+1} \mid X_k\right) b_{k+2:T}$$

- The time and space needed for each backward message are constant.
- Thus, the process of smoothing with respect to $e_{1:T}$ is *O(t)*.
- Thus, to smooth the whole sequence naively, requires $O(t^2)$.
- using dynamic programming the cost is only *O(t)* by recording results of forward filtering over the entire sequence while running the backward algorithm from *T* to 1 and use the smoothed message at each time step ➔ **forward-backward algo**.
    - space is now $O\left(\left|f\right|t\right)$
- In on-line setting, smoothed estimates must be computed for earlier time slices as new observations are added:
    - **fixed-lag smoothing –** smoothing is done for the time slice *d* steps behind the current time *T*.

Most Likely Explanation – task of finding the sequence of states most likely to have generated a sequence of observations; $\arg\max_{x_{1:t}} P(x_{1:t} \mid e_{1:t})$.

- most likely sequence must consider joint probabilities over all time steps.
- *there is a recursive relationship between most likely paths to each state X_{t+1} and the most likely paths to each state X_t.*
- Recursive formulation:

$$\max_{X_{1:t-1}} P(X_{1:t} \mid e_{1:t}) \propto \underbrace{P(e_t \mid X_t)}_{observation} \max_{X_{t-1}} \left[\underbrace{P(X_t \mid X_{t-1})}_{transition} \underbrace{\max_{X_{1:t-2}} P(X_{1:t-1} \mid e_{1:t-1})}_{\text{previous message}} \right]$$

  - messages: $m_{1:t} = \max_{X_{1:t-1}} P(X_{1:t} \mid e_{1:t})$
  - summation over X_t replaced by a maximization.
- Pointers are used to retrieve the most-likely explanation
- Viterbi algorithm has a space and time requirement of *O(t)*.

Learning – task of learning the transition and sensor models from observed data. This process leverages inference through EM.

Hidden Markov Models (HMM) – a temporal probabilistic model in which the state of the process is described by a *single discrete* random variable and transitions obey the Markov assumption.

- transition model: $T_{ij} = P(X_t = j \mid X_{t-1} = i)$
- observation model: $(\mathbf{O_t})_{i,i} = P(e_t \mid X_t = i)$
  - *forward* message - $\mathbf{f}_{1:t+1} \propto \mathbf{O}_{t+1}\mathbf{T}^T\mathbf{f}_{1:t}$
  - *backward* message - $\mathbf{b}_{k+1:t} \propto \mathbf{T}\mathbf{O}_{k+1}\mathbf{b}_{k+2:t}$
  - time complexity of forward-backward becomes $O(S^2 t)$ where S is the number of hidden states and space complexity is $O(St)$.

Kalman Filters – a temporal probabilistic model for continuous state spaces under the Markov assumption and using linear Gaussian distributions to model the states. A Kalman filter can model any system of continuous state variables with noisy measurements.

- a *multivariate Gaussian* distribution can be specified completely by its mean $\boldsymbol{\mu}$ and its covariance matrix $\boldsymbol{\Sigma}$.
- In general, filtering with continuous or hybrid spaces generate state distributions whose representations grow without bound, but the Gaussian distribution is "well-behaved" since it has the following properties:
  1. If the current distribution $P(\mathbf{X}_t \mid \mathbf{e}_{1:t})$ is Gaussian and the transition model $P(\mathbf{X}_{t+1} \mid \mathbf{x}_t)$ is linear Gaussian, then the predicted distribution of the next step is:
  $$P(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) = \int_{\mathbf{x}_t} P(\mathbf{X}_{t+1} \mid \mathbf{x}_t) P(\mathbf{x}_t \mid \mathbf{e}_{1:t}) d\mathbf{x}_t$$
  2. If the predicted distribution is Gaussian and the observation (sensor) model is linear Gaussian, then conditioning on new evidence yields the updated distribution:
  $$P(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) \propto P(\mathbf{e}_{1:t+1} \mid \mathbf{X}_{t+1}) P(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t})$$
- General formulation:
  $$P(\mathbf{x}_{t+1} \mid \mathbf{x}_t) = N(\mathbf{F}\mathbf{x}_t, \boldsymbol{\Sigma}_x)(\mathbf{x}_{t+1})$$
  - $\mathbf{F}$ and $\boldsymbol{\Sigma}_x$ describe the linear transition model & noise.

  $$P(\mathbf{z}_t \mid \mathbf{x}_t) = N(\mathbf{H}\mathbf{x}_t, \boldsymbol{\Sigma}_z)(\mathbf{z}_t)$$
  - $\mathbf{H}$ and $\boldsymbol{\Sigma}_z$ describe the linear sensor model & noise.
- Updates:
  $$\boldsymbol{\mu}_{t+1} = \mathbf{F}\boldsymbol{\mu}_t + \mathbf{K}_{t+1}(\mathbf{z}_{t+1} - \mathbf{H}\mathbf{F}\boldsymbol{\mu}_t)$$
  $$\boldsymbol{\Sigma}_{t+1} = (\mathbf{I} - \mathbf{K}_{t+1})(\mathbf{F}\boldsymbol{\Sigma}_t\mathbf{F}^T + \boldsymbol{\Sigma}_x)$$
  - Kalman gain $K_{t+1} = (\mathbf{F}\boldsymbol{\Sigma}_t\mathbf{F}^T + \boldsymbol{\Sigma}_x)\mathbf{H}^T\left(\mathbf{H}(\mathbf{F}\boldsymbol{\Sigma}_t\mathbf{F}^T + \boldsymbol{\Sigma}_x)\mathbf{H}^T + \boldsymbol{\Sigma}_z\right)^{-1}$
    - A measure of "how seriously to take the new observation" relative to the prediction.
  - predicted state at t+1 is $\mathbf{F}\boldsymbol{\mu}_t$, predicted observation is $\mathbf{H}\mathbf{F}\boldsymbol{\mu}_t$, and error of predicted observation is $(\mathbf{z}_{t+1} - \mathbf{H}\mathbf{F}\boldsymbol{\mu}_t)$.
- Extended Kalman Filter (EKF) – allows for limited nonlinearity in the model by modeling the system *locally* as linear in $\mathbf{x}_t$ in the region of $\mathbf{x}_t = \boldsymbol{\mu}_t$.