

Outline

- ◇ Time and uncertainty
- ◇ Inference: filtering, prediction, smoothing
- ◇ Hidden Markov models
- ◇ Kalman filters (a brief mention)
- ◇ Dynamic Bayesian networks
- ◇ Particle filtering

Time and uncertainty

The world changes; we need to track and predict it

Diabetes management vs vehicle diagnosis

Basic idea: copy state and evidence variables for each time step

$\mathbf{X}_t$ = set of unobservable state variables at time t
e.g., *BloodSugar*, *StomachContents* _{t} , etc.

$\mathbf{E}_t$ = set of observable evidence variables at time t
e.g., *MeasuredBloodSugar*, *PulseRate* _{t} , *FoodEaten* _{t}

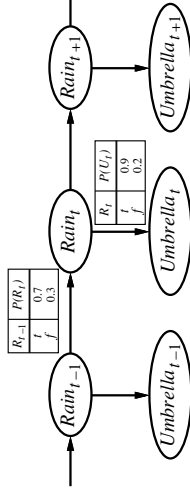
This assumes **discrete time**; step size depends on problem

Notation: $\mathbf{X}_{a:b} = \mathbf{X}_a, \mathbf{X}_{a+1}, \dots, \mathbf{X}_{b-1}, \mathbf{X}_b$

Sensor Markov assumption: $\mathbf{P}(\mathbf{E}_t | \mathbf{X}_{0:t}, \mathbf{E}_{0:t-1}) = \mathbf{P}(\mathbf{E}_t | \mathbf{X}_t)$

Stationary process: transition model $\mathbf{P}(\mathbf{X}_t | \mathbf{X}_{t-1})$ and sensor model $\mathbf{P}(\mathbf{E}_t | \mathbf{X}_t)$ fixed for all t

Example



First-order Markov assumption not exactly true in real world!

Possible fixes:

1. **Increase order** of Markov process
2. **Augment state**, e.g., add *Temp* _{t} , *Pressure* _{t}

Example: robot motion.

Augment position and velocity with *Battery* _{t}

Inference tasks

Filtering: $\mathbf{P}(\mathbf{X}_t | \mathbf{e}_{1:t})$
belief state—input to the decision process of a rational agent

Prediction: $\mathbf{P}(\mathbf{X}_{t+k} | \mathbf{e}_{1:t})$ for $k > 0$

evaluation of possible action sequences;
like filtering without the evidence

Smoothing: $\mathbf{P}(\mathbf{X}_k | \mathbf{e}_{1:t})$ for $0 \leq k < t$

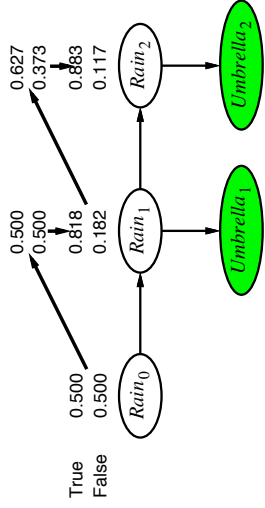
better estimate of past states, essential for learning

Most likely explanation: $\arg \max_{\mathbf{x}_{1:t}} P(\mathbf{x}_{1:t} | \mathbf{e}_{1:t})$

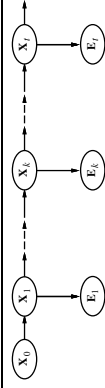
speech recognition, decoding with a noisy channel

$\mathbf{f}_{1:t+1} = \text{FORWARD}(\mathbf{f}_{1:t}, \mathbf{e}_{t+1})$ where $\mathbf{f}_{1:t} = \mathbf{P}(\mathbf{X}_t | \mathbf{e}_{1:t})$
Time and space **constant** (independent of t)

Filtering example



Smothing



Divide evidence $\mathbf{e}_{1:t}$ into $\mathbf{e}_{1:k}$, $\mathbf{e}_{k+1:t}$:

$$\begin{aligned} \mathbf{P}(\mathbf{X}_k | \mathbf{e}_{1:t}) &= \mathbf{P}(\mathbf{X}_k | \mathbf{e}_{1:k}, \mathbf{e}_{k+1:t}) \\ &= \alpha \mathbf{P}(\mathbf{X}_k | \mathbf{e}_{1:k}) \mathbf{P}(\mathbf{e}_{k+1:t} | \mathbf{X}_k, \mathbf{e}_{1:k}) \\ &= \alpha \mathbf{P}(\mathbf{X}_k | \mathbf{e}_{1:k}) \mathbf{P}(\mathbf{e}_{k+1:t} | \mathbf{X}_k) \\ &= \alpha \mathbf{f}_{1:k} \mathbf{b}_{k+1:t} \end{aligned}$$

Backward message computed by a backwards recursion:

$$\begin{aligned} \mathbf{P}(\mathbf{e}_{k+1:t} | \mathbf{X}_k) &= \sum_{\mathbf{x}_{k+1}} \mathbf{P}(\mathbf{e}_{k+1:t} | \mathbf{X}_k, \mathbf{x}_{k+1}) \mathbf{P}(\mathbf{x}_{k+1} | \mathbf{X}_k) \\ &= \sum_{\mathbf{x}_{k+1}} \mathbf{P}(\mathbf{e}_{k+1:t} | \mathbf{x}_{k+1}) \mathbf{P}(\mathbf{x}_{k+1} | \mathbf{X}_k) \\ &= \sum_{\mathbf{x}_{k+1}} \mathbf{P}(\mathbf{e}_{k+1} | \mathbf{x}_{k+1}) \mathbf{P}(\mathbf{e}_{k+2:t} | \mathbf{x}_{k+1}) \mathbf{P}(\mathbf{x}_{k+1} | \mathbf{X}_k) \end{aligned}$$

Forward-backward algorithm: cache forward messages along the way
Time linear in t (polytree inference), space $O(d\mathbf{f})$

Most likely explanation

Most likely sequence $\neq$ sequence of most likely states!!!

Most likely path to each $\mathbf{x}_{t+1}$

= most likely path to **some** $\mathbf{x}_t$ plus one more step

$$\begin{aligned} \max_{\mathbf{x}_1, \dots, \mathbf{x}_t} \mathbf{P}(\mathbf{x}_1, \dots, \mathbf{x}_t, \mathbf{X}_{t+1} | \mathbf{e}_{1:t+1}) \\ = \mathbf{P}(\mathbf{e}_{t+1} | \mathbf{X}_{t+1}) \max_{\mathbf{x}_t} (\mathbf{P}(\mathbf{X}_{t+1} | \mathbf{x}_t) \max_{\mathbf{x}_1, \dots, \mathbf{x}_{t-1}} \mathbf{P}(\mathbf{x}_1, \dots, \mathbf{x}_{t-1}, \mathbf{x}_t | \mathbf{e}_{1:t})) \end{aligned}$$

Identical to filtering, except $\mathbf{f}_{1:t}$ replaced by

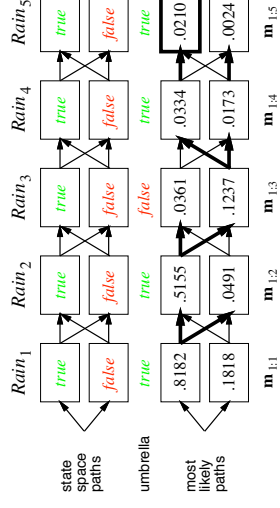
$$\mathbf{m}_{1:t} = \max_{\mathbf{x}_1, \dots, \mathbf{x}_{t-1}} \mathbf{P}(\mathbf{x}_1, \dots, \mathbf{x}_{t-1}, \mathbf{X}_t | \mathbf{e}_{1:t}),$$

i.e., $\mathbf{m}_{1:t}(i)$ gives the probability of the most likely path to state i .

Update has sum replaced by max, giving the **Viterbi algorithm**:

$$\mathbf{m}_{1:t+1} = \mathbf{P}(\mathbf{e}_{t+1} | \mathbf{X}_{t+1}) \max_{\mathbf{x}_t} (\mathbf{P}(\mathbf{X}_{t+1} | \mathbf{x}_t) \mathbf{m}_{1:t})$$

Viterbi example

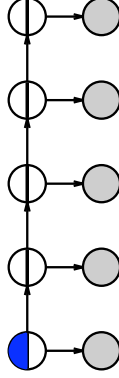


Country dance algorithm

Can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned}\mathbf{f}_{i,t+1} &= \alpha \mathbf{O}_{i+1} \mathbf{T}^\top \mathbf{f}_{i,t} \\ \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \alpha \mathbf{T}^\top \mathbf{f}_{i,t} \\ \alpha' (\mathbf{T}^\top)^{-1} \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \mathbf{f}_{i,t}\end{aligned}$$

Algorithm: forward pass computes $\mathbf{f}_i$, backward pass does $\mathbf{f}_i, \mathbf{b}_i$

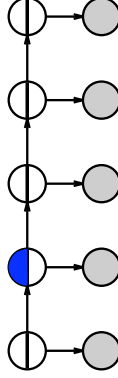


Country dance algorithm

Can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned}\mathbf{f}_{i,t+1} &= \alpha \mathbf{O}_{i+1} \mathbf{T}^\top \mathbf{f}_{i,t} \\ \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \alpha \mathbf{T}^\top \mathbf{f}_{i,t} \\ \alpha' (\mathbf{T}^\top)^{-1} \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \mathbf{f}_{i,t}\end{aligned}$$

Algorithm: forward pass computes $\mathbf{f}_i$, backward pass does $\mathbf{f}_i, \mathbf{b}_i$

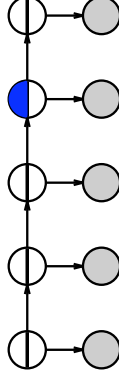


Country dance algorithm

Can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned}\mathbf{f}_{i,t+1} &= \alpha \mathbf{O}_{i+1} \mathbf{T}^\top \mathbf{f}_{i,t} \\ \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \alpha \mathbf{T}^\top \mathbf{f}_{i,t} \\ \alpha' (\mathbf{T}^\top)^{-1} \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \mathbf{f}_{i,t}\end{aligned}$$

Algorithm: forward pass computes $\mathbf{f}_i$, backward pass does $\mathbf{f}_i, \mathbf{b}_i$

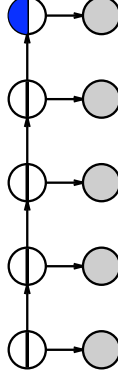


Country dance algorithm

Can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned}\mathbf{f}_{i,t+1} &= \alpha \mathbf{O}_{i+1} \mathbf{T}^\top \mathbf{f}_{i,t} \\ \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \alpha \mathbf{T}^\top \mathbf{f}_{i,t} \\ \alpha' (\mathbf{T}^\top)^{-1} \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i,t+1} &= \mathbf{f}_{i,t}\end{aligned}$$

Algorithm: forward pass computes $\mathbf{f}_i$, backward pass does $\mathbf{f}_i, \mathbf{b}_i$

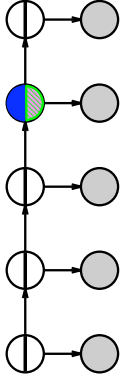


Country dance algorithm

Can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned}\mathbf{f}_{i:t+1} &= \alpha \mathbf{O}_{i+1} \mathbf{T}^T \mathbf{f}_{i:t} \\ \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i:t+1} &= \alpha \mathbf{T}^T \mathbf{f}_{i:t} \\ \alpha' (\mathbf{T}^T)^{-1} \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i:t+1} &= \mathbf{f}_{i:t}\end{aligned}$$

Algorithm: forward pass computes $\mathbf{f}_i$, backward pass does $\mathbf{f}_i$, $\mathbf{b}_i$

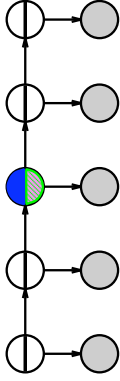


Country dance algorithm

Can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned}\mathbf{f}_{i:t+1} &= \alpha \mathbf{O}_{i+1} \mathbf{T}^T \mathbf{f}_{i:t} \\ \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i:t+1} &= \alpha \mathbf{T}^T \mathbf{f}_{i:t} \\ \alpha' (\mathbf{T}^T)^{-1} \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i:t+1} &= \mathbf{f}_{i:t}\end{aligned}$$

Algorithm: forward pass computes $\mathbf{f}_i$, backward pass does $\mathbf{f}_i$, $\mathbf{b}_i$

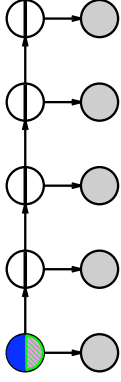


Country dance algorithm

Can avoid storing all forward messages in smoothing by running forward algorithm backwards:

$$\begin{aligned}\mathbf{f}_{i:t+1} &= \alpha \mathbf{O}_{i+1} \mathbf{T}^T \mathbf{f}_{i:t} \\ \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i:t+1} &= \alpha \mathbf{T}^T \mathbf{f}_{i:t} \\ \alpha' (\mathbf{T}^T)^{-1} \mathbf{O}_{i+1}^{-1} \mathbf{f}_{i:t+1} &= \mathbf{f}_{i:t}\end{aligned}$$

Algorithm: forward pass computes $\mathbf{f}_i$, backward pass does $\mathbf{f}_i$, $\mathbf{b}_i$

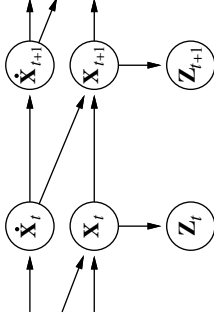


Kalman filters

Modelling systems described by a set of continuous variables,

e.g., tracking a bird flying— $\mathbf{X}_t = X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}$.

Airplanes, robots, ecosystems, economies, chemical plants, planets, ...



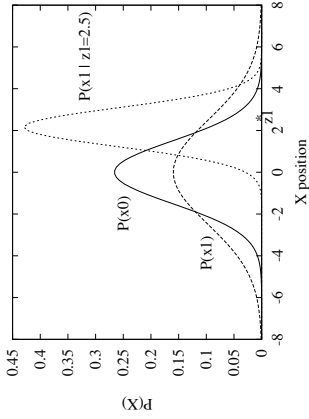
Gaussian prior, linear Gaussian transition model and sensor model

Simple 1-D example

Gaussian random walk on X -axis, s.d. σ_x , sensor s.d. σ_z

$$\mu_{t+1} = \frac{(\sigma_t^2 + \sigma_x^2)z_{t+1} + \sigma_z^2\mu_t}{\sigma_t^2 + \sigma_x^2 + \sigma_z^2}$$

$$\sigma_{t+1}^2 = \frac{(\sigma_t^2 + \sigma_x^2)\sigma_z^2}{\sigma_t^2 + \sigma_x^2 + \sigma_z^2}$$



General Kalman update

Transition and sensor models:

$$P(\mathbf{x}_{t+1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{F}\mathbf{x}_t, \Sigma_x)(\mathbf{x}_{t+1})$$

$$P(z_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{H}\mathbf{x}_t, \Sigma_z)(z_t)$$

$\mathbf{F}$ is the matrix for the transition; Σ_x the transition noise covariance

$\mathbf{H}$ is the matrix for the sensors; Σ_z the sensor noise covariance

Filter computes the following update:

$$\mu_{t+1} = \mathbf{F}\mu_t + \mathbf{K}_{t+1}(z_{t+1} - \mathbf{H}\mathbf{F}\mu_t)$$

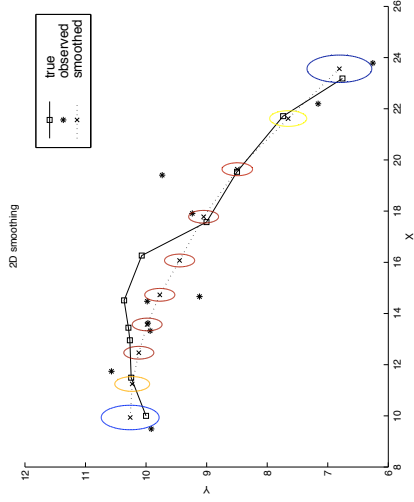
$$\Sigma_{t+1} = (\mathbf{I} - \mathbf{K}_{t+1})(\mathbf{F}\Sigma_t\mathbf{F}^T + \Sigma_x)$$

where $\mathbf{K}_{t+1} = (\mathbf{F}\Sigma_t\mathbf{F}^T + \Sigma_x)\mathbf{H}^T(\mathbf{H}(\mathbf{F}\Sigma_t\mathbf{F}^T + \Sigma_x)\mathbf{H}^T + \Sigma_z)^{-1}$ is the Kalman gain matrix

Σ_t and $\mathbf{K}_t$ are independent of observation sequence, so compute offline



2-D tracking example: smoothing

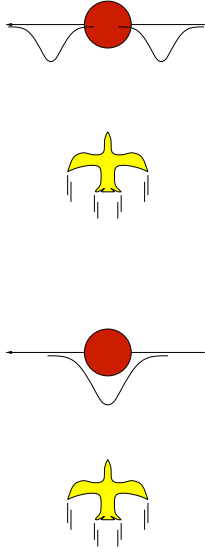


Where it breaks

Cannot be applied if the transition model is nonlinear

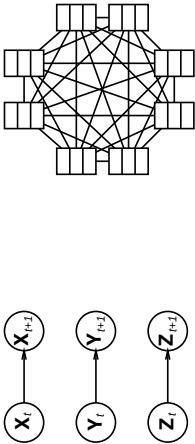
Extended Kalman Filter models transition as **locally linear** around $\mathbf{x}_t = \mu_t$

Fails if systems is locally unsmooth



DBNs vs. HMMs

Every HMM is a single-variable DBN; every discrete DBN is an HMM

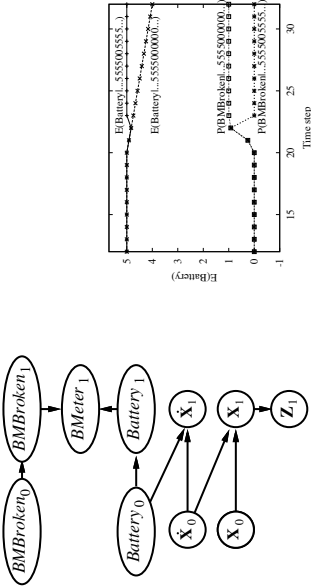


Sparse dependencies $\Rightarrow$ exponentially fewer parameters;
e.g., 20 state variables, three parents each
DBN has $20 \times 2^3 = 160$ parameters, HMM has $2^{20} \times 2^{20} \approx 10^{12}$

DBNs vs Kalman filters

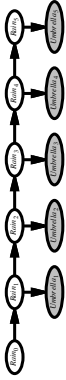
Every Kalman filter model is a DBN, but few DBNs are KFs;
real world requires non-Gaussian posteriors

E.g., where are bin Laden and my keys? What's the battery charge?

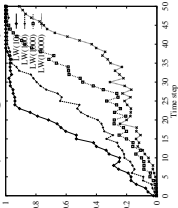


Likelihood weighting for DBNs

Set of weighted samples approximates the belief state



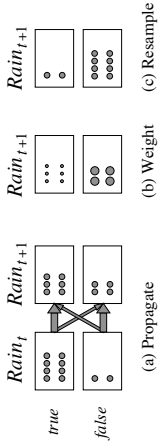
LW samples pay no attention to the evidence!
 $\Rightarrow$ fraction "agreeing" falls exponentially with t
 $\Rightarrow$ number of samples required grows exponentially with t



Particle filtering

Basic idea: ensure that the population of samples ("particles") tracks the high-likelihood regions of the state-space

Replicate particles proportional to likelihood for e_t

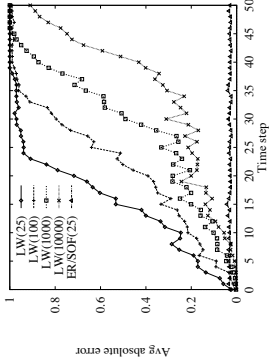


Widely used for tracking nonlinear systems, esp. in vision
Also used for simultaneous localization and mapping in mobile robots
 10^5 -dimensional state space

$$= P(\mathbf{x}_{t+1}|\mathbf{e}_{1:t+1})$$

Particle filtering performance

Approximation error of particle filtering remains bounded over time, at least empirically—theoretical analysis is difficult



Summary

Temporal models use state and sensor variables replicated over time

- Markov assumptions and stationarity assumption, so we need
- transition model $P(\mathbf{X}_t|\mathbf{X}_{t-1})$
 - sensor model $P(\mathbf{E}_t|\mathbf{X}_t)$

Tasks are filtering, prediction, smoothing, most likely sequence;
all done recursively with constant cost per time step

Hidden Markov models have a single discrete state variable; used for speech recognition

Kalman filters allow n state variables, linear Gaussian, $O(n^3)$ update

Dynamic Bayes nets subsume HMMs, Kalman filters; exact update intractable

Particle filtering is a good approximate filtering algorithm for DBNs