Temporal probability models

Chapter 15, Sections 1–5
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

\[ X_t = \text{set of unobservable state variables at time } t \]

\[ \text{e.g., } BloodSugar_t, StomachContents_t, \text{ etc.} \]

\[ E_t = \text{set of observable evidence variables at time } t \]

\[ \text{e.g., } MeasuredBloodSugar_t, PulseRate_t, FoodEaten_t \]

This assumes discrete time; step size depends on problem

Notation: \[ X_{a:b} = X_a, X_{a+1}, \ldots, X_{b-1}, X_b \]
Markov processes (Markov chains)

Construct a Bayes net from these variables: parents?

**Markov assumption:** $X_t$ depends on bounded subset of $X_{0:t-1}$

**First-order Markov process:** $P(X_t|X_{0:t-1}) = P(X_t|X_{t-1})$

**Second-order Markov process:** $P(X_t|X_{0:t-1}) = P(X_t|X_{t-2}, X_{t-1})$

**Sensor Markov assumption:** $P(E_t|X_{0:t}, E_{0:t-1}) = P(E_t|X_t)$

**Stationary process:** transition model $P(X_t|X_{t-1})$ and sensor model $P(E_t|X_t)$ fixed for all $t$
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: $\arg\max_{\mathbf{x}_{1:t}} \mathbf{P}(\mathbf{x}_{1:t}|\mathbf{e}_{1:t})$

speech recognition, decoding with a noisy channel
Filtering

Aim: devise a **recursive** state estimation algorithm:

\[ P(X_{t+1}|e_{1:t+1}) = f(e_{t+1}, P(X_t|e_{1:t})) \]

\[
P(X_{t+1}|e_{1:t+1}) = P(X_{t+1}|e_{1:t}, e_{t+1})
= \alpha P(e_{t+1}|X_{t+1}, e_{1:t})P(X_{t+1}|e_{1:t})
= \alpha P(e_{t+1}|X_{t+1})P(X_{t+1}|e_{1:t})
\]

I.e., prediction + estimation. Prediction by summing out \( X_t \):

\[
P(X_{t+1}|e_{1:t+1}) = \alpha P(e_{t+1}|X_{t+1})\sum_{x_t}P(X_{t+1}|x_t, e_{1:t})P(x_t|e_{1:t})
= \alpha P(e_{t+1}|X_{t+1})\sum_{x_t}P(X_{t+1}|x_t)P(x_t|e_{1:t})
\]

\[ f_{1:t+1} = \text{FORWARD}(f_{1:t}, e_{t+1}) \text{ where } f_{1:t} = P(X_t|e_{1:t}) \]

Time and space **constant** (independent of \( t \)) for finite-state \( X_t \)
Filtering example

\begin{align*}
\text{True} & : \frac{1}{2} \\
\text{False} & : \frac{1}{2}
\end{align*}

\begin{align*}
\text{Rain}_0 & \rightarrow \text{Rain}_1 \\
\text{Rain}_1 & \rightarrow \text{Rain}_2 \\
\text{Rain}_0 & \rightarrow \text{Umbrella}_1 \\
\text{Rain}_2 & \rightarrow \text{Umbrella}_2
\end{align*}
Divide evidence $e_{1:t}$ into $e_{1:k}$, $e_{k+1:t}$:

$$P(X_k|e_{1:t}) = P(X_k|e_{1:k}, e_{k+1:t})$$
$$= \alpha P(X_k|e_{1:k}) P(e_{k+1:t}|X_k, e_{1:k})$$
$$= \alpha P(X_k|e_{1:k}) P(e_{k+1:t}|X_k)$$
$$= \alpha f_{1:k} b_{k+1:t}$$

Backward message computed by a backwards recursion:

$$P(e_{k+1:t}|X_k) = \sum_{x_{k+1}} P(e_{k+1:t}|X_k, x_{k+1}) P(x_{k+1}|X_k)$$
$$= \sum_{x_{k+1}} P(e_{k+1:t}|x_{k+1}) P(x_{k+1}|X_k)$$
$$= \sum_{x_{k+1}} P(e_{k+1}|x_{k+1}) P(e_{k+2:t}|x_{k+1}) P(x_{k+1}|X_k)$$
Smoothing example

Forward–backward algorithm: cache forward messages along the way
Time linear in $t$ (polytree inference), space $O(t|f|)$
Most likely explanation

Most likely sequence ≠ sequence of most likely states!!!!

Most likely path to each $x_{t+1}$

= most likely path to some $x_t$ plus one more step

$$\max_{x_1 \ldots x_t} P(x_1, \ldots, x_t, X_{t+1} | e_{1:t+1})$$

$$= P(e_{t+1} | X_{t+1}) \max_{x_t} \left( P(X_{t+1} | x_t) \max_{x_1 \ldots x_{t-1}} P(x_1, \ldots, x_{t-1}, x_t | e_{1:t}) \right)$$

Identical to filtering, except $f_{1:t}$ replaced by

$$\mathbf{m}_{1:t} = \max_{x_1 \ldots x_{t-1}} \mathbf{P}(x_1, \ldots, x_{t-1}, X_t | e_{1:t}),$$

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:

$$\mathbf{m}_{1:t+1} = P(e_{t+1} | X_{t+1}) \max_{x_t} (P(X_{t+1} | x_t) \mathbf{m}_{1:t})$$
Viterbi example

\begin{center}
\begin{tabular}{cccccc}
\text{Rain}_1 & \text{Rain}_2 & \text{Rain}_3 & \text{Rain}_4 & \text{Rain}_5 \\
true & true & true & true & true \\
false & false & false & false & false \\
\end{tabular}
\end{center}

\begin{center}
\begin{tabular}{cccccc}
true & true & false & true & true \\
.8182 & .5155 & .0361 & .0334 & .0210 \\
.1818 & .0491 & .1237 & .0173 & .0024 \\
\end{tabular}
\end{center}

\begin{center}
\text{state} \quad \text{space} \quad \text{paths} \\
\text{umbrella} \quad \text{true} \quad \text{false} \\
\text{most} \quad \text{likely} \quad \text{paths} \\
\text{m}_{1:1} \quad \text{m}_{1:2} \quad \text{m}_{1:3} \quad \text{m}_{1:4} \quad \text{m}_{1:5}
\end{center}
**Hidden Markov models**

\( X_t \) is a single, discrete variable (usually \( E_t \) is too)

Domain of \( X_t \) is \( \{1, \ldots, S\} \)

Transition matrix \( T_{ij} = P(X_t = j | X_{t-1} = i) \), e.g., \( \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix} \)

Sensor matrix \( O_t \) for each time step, diagonal elements \( P(e_t | X_t = i) \)

E.g., with \( U_1 = true \), \( O_1 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.2 \end{pmatrix} \)

Forward and backward messages as column vectors:

\[
\begin{align*}
  f_{1:t+1} &= \alpha O_{t+1} T^T f_{1:t} \\
  b_{k+1:t} &= T O_{k+1} b_{k+2:t}
\end{align*}
\]

Forward–backward algorithm needs time \( O(S^2 t) \) and space \( O(St) \)
Country dance algorithm

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

\[
\begin{align*}
f_{1:t+1} &= \alpha O_{t+1} T^\top f_{1:t} \\
O_{t+1}^{-1} f_{1:t+1} &= \alpha T^\top f_{1:t} \\
\alpha'(T^\top)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[ f_{1:t+1} = \alpha O_{t+1} T^\top f_{1:t} \]
\[ O_{t+1}^{-1} f_{1:t+1} = \alpha T^\top f_{1:t} \]
\[ \alpha'(T^\top)^{-1} O_{t+1}^{-1} f_{1:t+1} = f_{1:t} \]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[
\begin{align*}
    f_{1:t+1} &= \alpha O_{t+1} T^\top f_{1:t} \\
    O_{t+1}^{-1} f_{1:t+1} &= \alpha T^\top f_{1:t} \\
    \alpha'(T^\top)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[
\begin{align*}
    f_{1:t+1} &= \alpha O_{t+1} T^\top f_{1:t} \\
    O_t^{-1} f_{1:t+1} &= \alpha T^\top f_{1:t} \\
    \alpha'(T^\top)^{-1} O_t^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[
\begin{align*}
    f_{1:t+1} &= \alpha O_{t+1} T^T f_{1:t} \\
    O_{t+1}^{-1} f_{1:t+1} &= \alpha T^T f_{1:t} \\
    \alpha'(T^T)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[
\begin{align*}
    f_{1:t+1} &= \alpha O_{t+1} T^\top f_{1:t} \\
    O_{t+1}^{-1} f_{1:t+1} &= \alpha T^\top f_{1:t} \\
    \alpha'(T^\top)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[
\begin{align*}
f_{1:t+1} &= \alpha O_{t+1} T^T f_{1:t} \\
O_{t+1}^{-1} f_{1:t+1} &= \alpha T^T f_{1:t} \\
\alpha'(T^T)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[
\begin{align*}
    f_{1:t+1} &= \alpha O_{t+1} T^\top f_{1:t} \\
    O_{t+1}^{-1} f_{1:t+1} &= \alpha T^\top f_{1:t} \\
    \alpha' (T^\top)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
**Country dance algorithm**

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

\[
\begin{align*}
    f_{1:t+1} &= \alpha O_{t+1} T^\top f_{1:t} \\
    O_{t+1}^{-1} f_{1:t+1} &= \alpha T^\top f_{1:t} \\
    \alpha'(T^\top)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \( f_t \), backward pass does \( f_i, b_i \)
Country dance algorithm

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

\[
\begin{align*}
  f_{1:t+1} &= \alpha O_{t+1} T^\top f_{1:t} \\
  O_{t+1}^{-1} f_{1:t+1} &= \alpha T^\top f_{1:t} \\
  \alpha'(T^\top)^{-1} O_{t+1}^{-1} f_{1:t+1} &= f_{1:t}
\end{align*}
\]

Algorithm: forward pass computes \(f_t\), backward pass does \(f_i, b_i\)
Kalman filters

Modelling systems described by a set of continuous variables, e.g., tracking a bird flying—\( X_t = X, Y, Z, \dot{X}, \dot{Y}, \dot{Z} \).

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

Gaussian prior, linear Gaussian transition model and sensor model
Predicting step: if $P(X_t|e_{1:t})$ is Gaussian, then prediction

$$P(X_{t+1}|e_{1:t}) = \int_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t}) \, dx_t$$

is Gaussian. If $P(X_{t+1}|e_{1:t})$ is Gaussian, then the updated distribution

$$P(X_{t+1}|e_{1:t+1}) = \alpha P(e_{t+1}|X_{t+1})P(X_{t+1}|e_{1:t})$$

is Gaussian.

Hence $P(X_t|e_{1:t})$ is multivariate Gaussian $\mathcal{N}(\mu_t, \Sigma_t)$ for all $t$.

General (nonlinear, non-Gaussian) process: description of posterior grows unboundedly as $t \to \infty$. 

Chapter 15, Sections 1–5  
25
Simple 1-D example

Gaussian random walk on $X$–axis, s.d. $\sigma_x$, sensor s.d. $\sigma_z$

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

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

Transition and sensor models:

\[ P(x_{t+1}|x_t) = N(Fx_t, \Sigma_x)(x_{t+1}) \]
\[ P(z_t|x_t) = N(Hx_t, \Sigma_z)(z_t) \]

\( F \) is the matrix for the transition; \( \Sigma_x \) the transition noise covariance
\( H \) is the matrix for the sensors; \( \Sigma_z \) the sensor noise covariance

Filter computes the following update:

\[ \mu_{t+1} = F\mu_t + K_{t+1}(z_{t+1} - HF\mu_t) \]
\[ \Sigma_{t+1} = (I - K_{t+1})(F\Sigma_tF^T + \Sigma_x) \]

where \( K_{t+1} = (F\Sigma_tF^T + \Sigma_x)H^T(H(F\Sigma_tF^T + \Sigma_x)H^T + \Sigma_z)^{-1} \)

is the Kalman gain matrix

\( \Sigma_t \) and \( K_t \) are independent of observation sequence, so compute offline
2-D tracking example: filtering
2-D tracking example: smoothing

![Diagram showing 2D smoothing with true, observed, and smoothed data points.]
Where it breaks

Cannot be applied if the transition model is nonlinear

Extended Kalman Filter models transition as locally linear around $x_t = \mu_t$

Fails if systems is locally unsmooth
$X_t, E_t$ contain arbitrarily many variables in a replicated Bayes net
DBNs vs. HMMs

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

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

Naive method: unroll the network and run any exact algorithm

Problem: inference cost for each update grows with $t$

Rollup filtering: add slice $t + 1$, “sum out” slice $t$ using variable elimination

Largest factor is $O(d^{n+1})$, update cost $O(d^{n+2})$
(cf. HMM update cost $O(d^{2n})$)
Likelihood weighting for DBNs

Set of weighted samples approximates the belief state

LW samples pay no attention to the evidence!

⇒ fraction “agreeing” falls exponentially with $t$
⇒ number of samples required grows exponentially with $t$

![Graph showing RMS error over time steps for LW samples with different sample sizes]
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 $e_t$

Widey used for tracking nonlinear systems, esp. in vision

Also used for simultaneous localization and mapping in mobile robots $10^5$-dimensional state space
Particle filtering contd.

Assume consistent at time $t$: $N(x_t|e_{1:t})/N = P(x_t|e_{1:t})$

Propagate forward: populations of $x_{t+1}$ are

$$N(x_{t+1}|e_{1:t}) = \sum_{x_t} P(x_{t+1}|x_t)N(x_t|e_{1:t})$$

Weight samples by their likelihood for $e_{t+1}$:

$$W(x_{t+1}|e_{1:t+1}) = P(e_{t+1}|x_{t+1})N(x_{t+1}|e_{1:t})$$

Resample to obtain populations proportional to $W$:

$$N(x_{t+1}|e_{1:t+1})/N = \alpha W(x_{t+1}|e_{1:t+1}) = \alpha P(e_{t+1}|x_{t+1})N(x_{t+1}|e_{1:t})$$

$$= \alpha P(e_{t+1}|x_{t+1})\sum_{x_t} P(x_{t+1}|x_t)N(x_t|e_{1:t})$$

$$= \alpha' P(e_{t+1}|x_{t+1})\sum_{x_t} P(x_{t+1}|x_t)P(x_t|e_{1:t})$$

$$= P(x_{t+1}|e_{1:t+1})$$
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 $P(X_t|X_{t-1})$
- sensor model $P(E_t|X_t)$

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