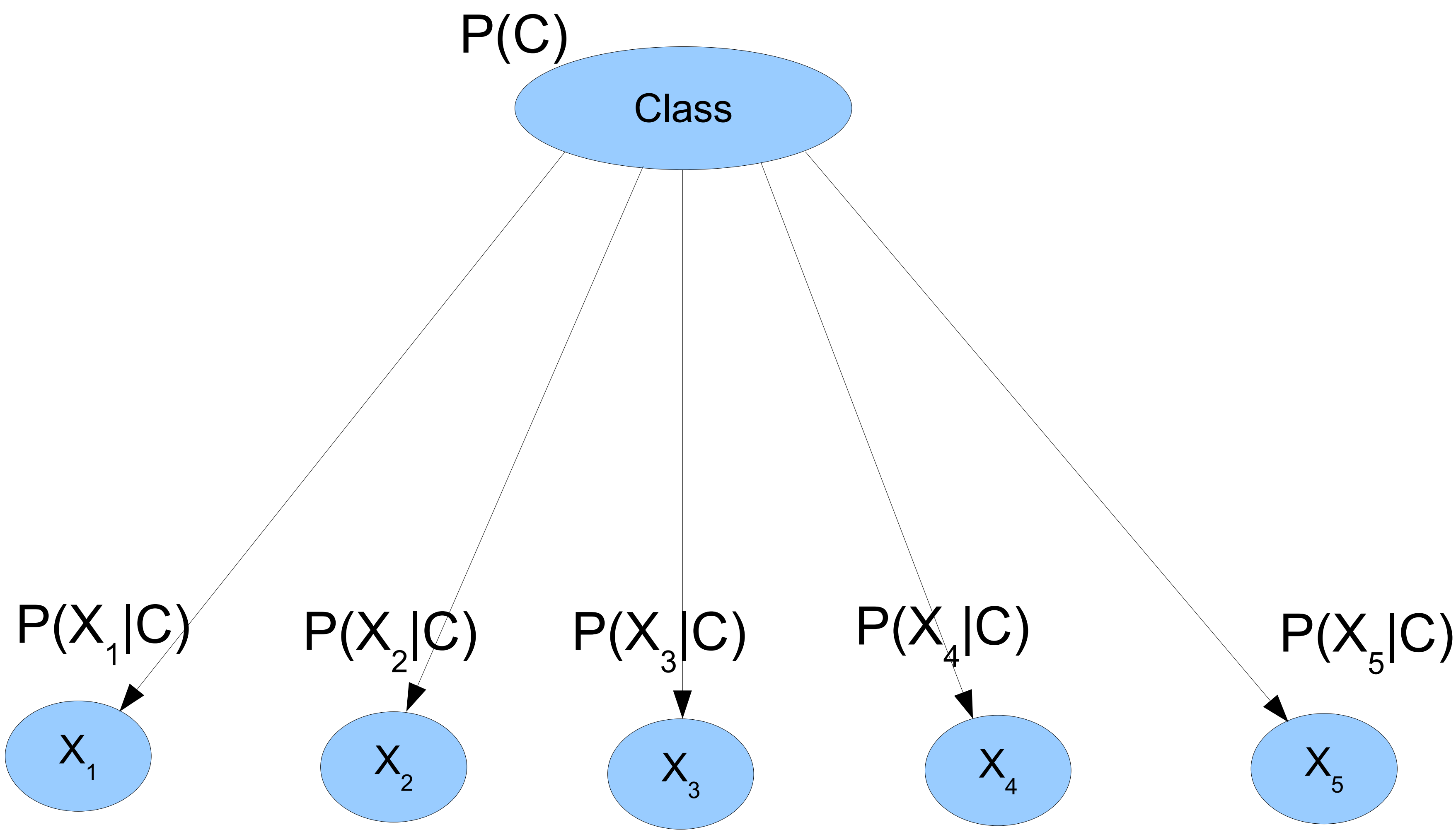
Some famous network models

- Naïve Bayes / Finite mixture model
- Tree Augmented Naïve Bayes
- Hidden Markov Models

Naïve Bayes classifier



- X_i are called predictors or indicators

Naïve Bayes Classifier

- Structure tailored for efficient diagnostics $P(C|x_1,x_2,...,x_n)$.
- Unrealistic conditional independence assumptions, but OK for the particular query $P(C|x_1,x_2,...,x_n)$.
- Because of wrong independence assumptions, NB is often poorly calibrated:
  - Probabilities $P(C|x_1,x_2,...,x_n)$ way off, but $\text{argmax}_c P(c|x_1,x_2,...,x_n)$ still often correct.

Calculating $P(C|x_1,x_2,\ldots,x_n,NB)$

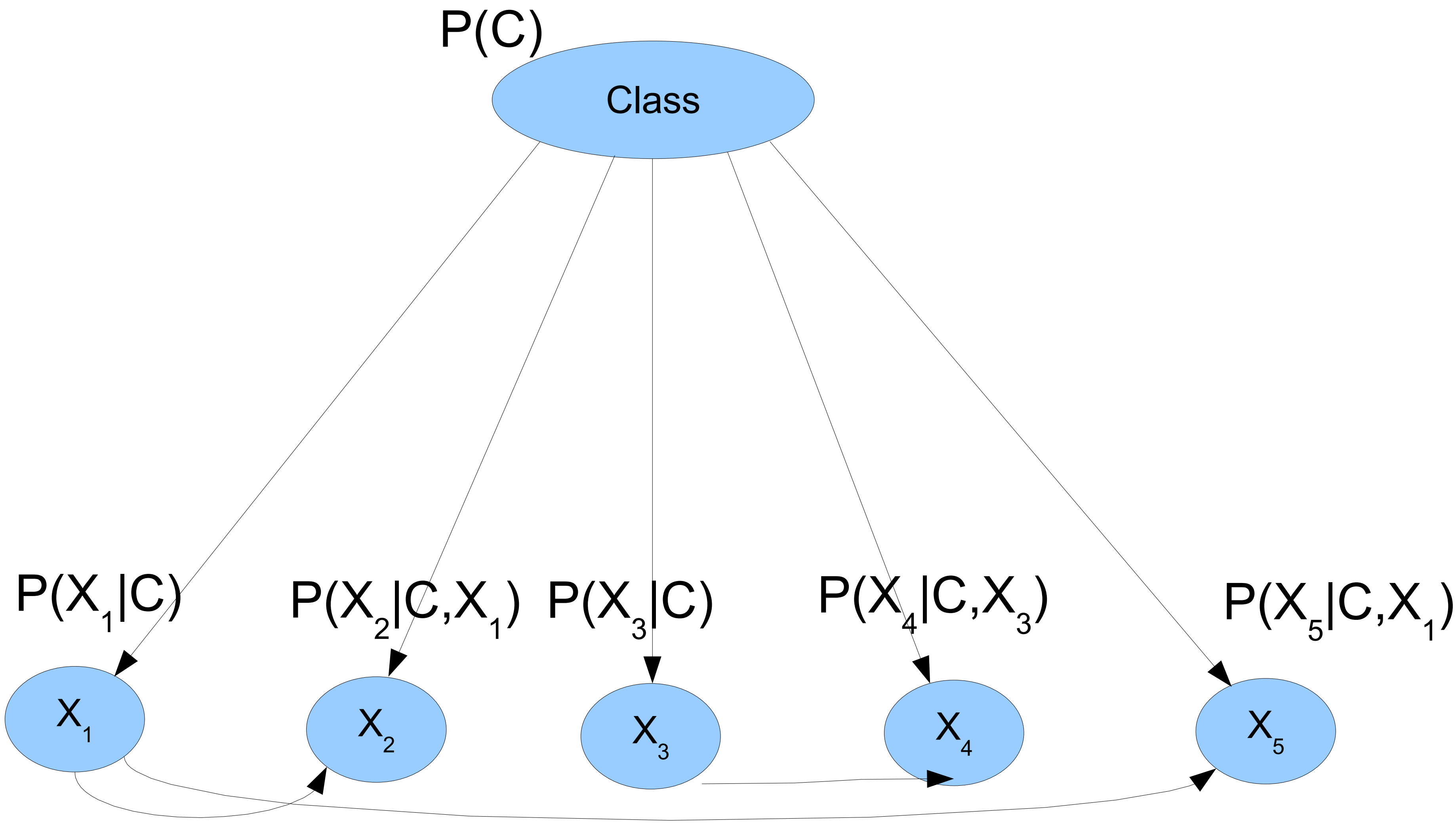
- Boldly calculate through joint probability

$$P(C|x_{1,}\ldots,x_n)\propto P(C,x_{1,}\ldots,x_n)=P(C)\prod_{i=1}^{n}P(x_i|C)$$

- No need to have all the predictors. Having just set X_A of predictors (and not X_B):

$$\begin{aligned}P(C|x_A)\propto P(C,x_A)&=\sum_{x_B}P(C,x_A,x_B)\\&=\sum_{x_B}P(C)\prod_{i\in A}P(x_i|C)\prod_{j\in B}P(x_j|C)\\&=P(C)\prod_{i\in A}P(x_i|C)\sum_{x_B}\prod_{j\in B}P(x_j|C)\\&=P(C)\prod_{i\in A}P(x_i|C)\prod_{j\in B}\sum_{x_j}P(x_j|C)=P(C)\prod_{i\in A}P(x_i|C)\end{aligned}$$

Tree Augmented Naïve Bayes (TAN)



- X_i may have at most one other X_j as an extra parent.

Calculating $P(C|x_1,x_2,\ldots,x_n,TAN)$

- Again, boldly calculate via joint probability

$$P(C|x_{1,},\ldots,x_n) \propto P(C,x_{1,},\ldots,x_n) = P(C)\prod_{i=1}^{n} P(x_i|C,Pa(x_i))$$

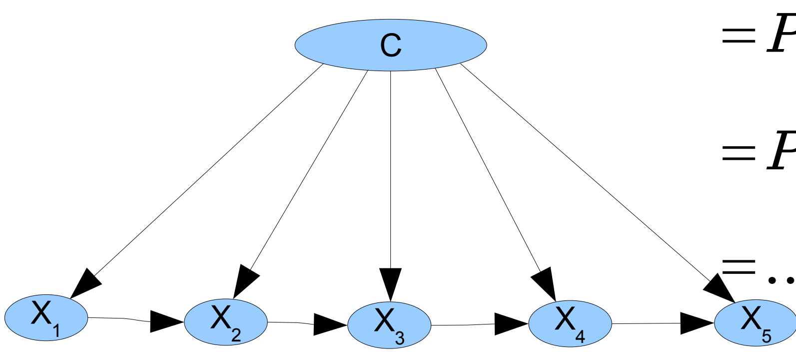
- But missing predictors may hurt more. For example, given the TAN in previous slide:

$$P(C|x_5) \propto P(C)P(x_5|C) = P(C)\sum_{x_4} P(x_4|C)P(x_5|x_{4,}C)$$

$$= P(C)\sum_{x_4} P(x_5|C,x_4)P(x_4|C)$$

$$= P(C)\sum_{x_4} P(x_5|C,x_4)\sum_{x_3} P(x_4|C,x_3)P(x_3|C)$$

$$= \ldots$$



NB as a Finite Mixture Model

- When NB structure is right, it also makes a nice (marginal) joint probability model $P(X_1,X_2,...,X_n)$ for “predictors”.
- A computationally effective alternative for building a Bayesian network for $X_1,X_2,...,X_n$.
- Joint probability $P(X_1,X_2,...,X_n)$ is represented as a mixture of K joint probability distributions $P_k(X_1,X_2,...,X_n) = P_k(X_1)P_k(X_2)...P_k(X_n)$, where $P_k(\cdot) = P(\cdot|C=k)$.

Calculating with $P(X_1, X_2, \ldots, X_n | NB)$

- Joint probability a simple marginalization:

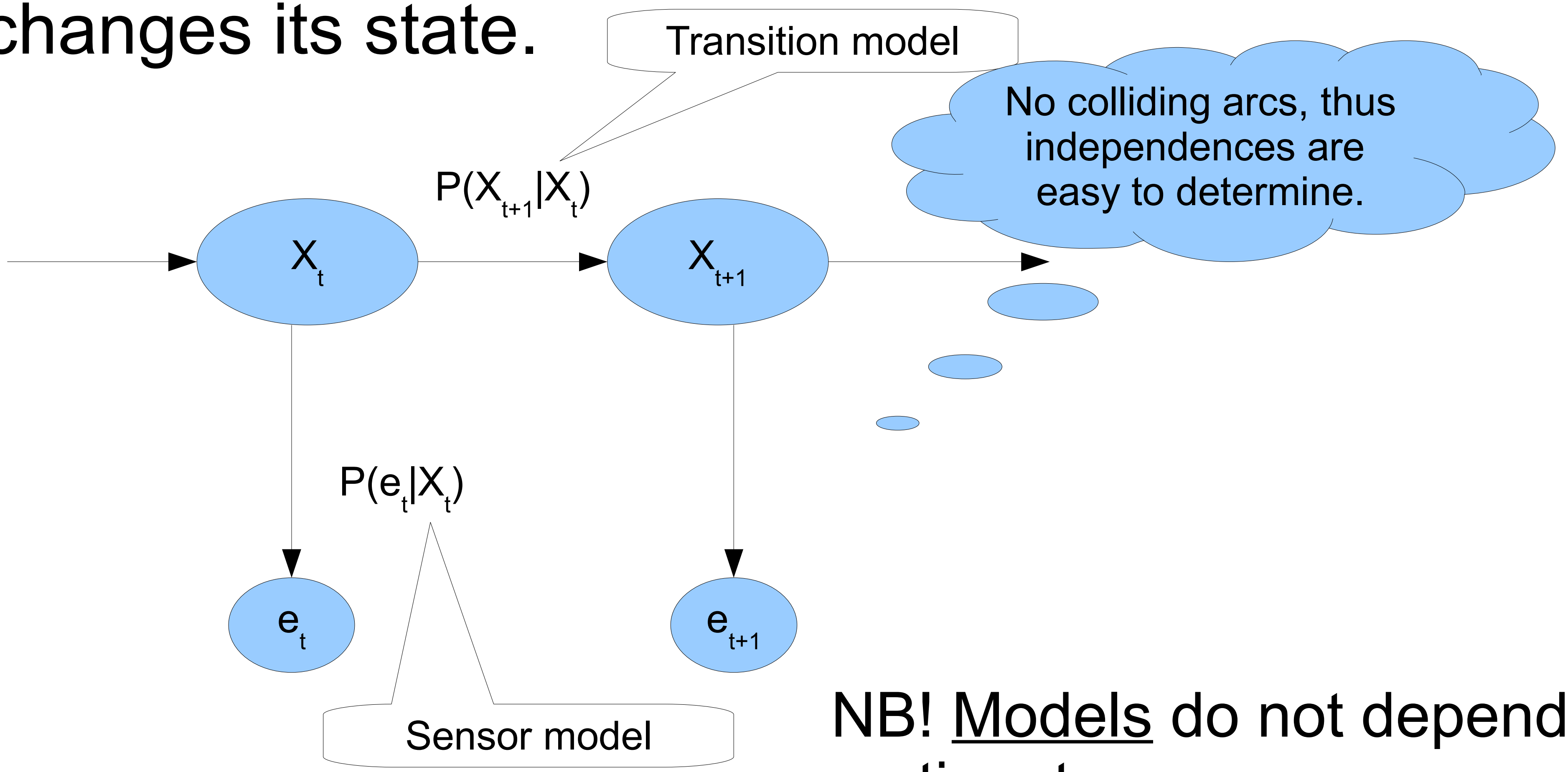
$$P(X_1, \ldots, X_n) = \sum_{k=1}^{K} P(X_1, \ldots, X_n, C=k)$$
$$= \sum_{k=1}^{K} P(C=k) \prod_{i=1}^{n} P(X_i | C=k)$$

- Inference

$$P(X|e) \propto P(e, X) = \sum_{k=1}^{K} P(e, X, C=k)$$
$$= \sum_{k=1}^{K} P(C=k) P(e, X | C=k)$$
$$= \sum_{k=1}^{K} \prod_{X_i \in X} P(X_i | C=k) \prod_{e_i \in e} P(e_i | C=k)$$

Hidden Markov Models

- Models observations about a system that changes its state.



NB! <u>Models</u> do not depend on time t.

Joint probability

- Joint probability like in Bayesian network
  - HMM is a Bayesian network

$$P(X_0, X_1, E_1, X_2, E_2, \ldots, X_t, E_t) = P(X_0) \prod_{i=1}^{t} P(X_i | X_{i-1}) P(E_i | X_i)$$

- Common inference tasks:
  - Filtering / monitoring: $P(X_t \mid e_{1:t})$
  - Prediction: $P(X_{t+k} \mid e_{1:t})$, $k>0$
  - Smoothing: $P(X_k \mid e_{1:t})$, $k<t$
  - Explanation: $P(X_{1:t} \mid e_{1:t})$

Calculating $P(X_t \mid e_{1:t})$ in HMM

- Lets shoot for a recursive formula:

$$
\begin{aligned}
P(X_{t+1} \mid e_{1:t+1}) &= P(X_{t+1} \mid e_{t+1}, e_{1:t}) \\
&\propto P(e_{t+1} \mid X_{t+1}, e_{1:t})\, P(X_{t+1} \mid e_{1:t}) \\
&= P(e_{t+1} \mid X_{t+1})\, \underline{P(X_{t+1} \mid e_{1:t})}
\end{aligned}
$$

- and

$$
\begin{aligned}
P(X_{t+1} \mid e_{1:t}) &= \sum_{x_t} P(X_{t+1}, x_t \mid e_{1:t}) \\
&= \sum_{x_t} P(X_{t+1} \mid x_t, e_{1:t})\, P(X_t \mid e_{1:t}) \\
&= \sum_{x_t} P(X_{t+1} \mid x_t)\, \underline{P(x_t \mid e_{1:t})}
\end{aligned}
$$

Forward algorithm for $P(X_t \mid e_{1:t})$

- Combining formulas we get a recursion

$$P(X_{t+1}|e_{1:t+1}) \propto P(e_{t+1}|X_{t+1}) \sum_{x_t} P(X_{t+1}|x_t) \underline{P(x_t|e_{1:t})}$$

- So first calculate

$$P(X_1|e_1) \propto P(e_1|X_1) \sum_{x_0} P(X_1|x_0) P(x_0)$$

- and then

$$P(X_2|e_{1,}e_2) \propto P(e_2|X_2) \sum_{x_1} P(X_2|x_1) P(x_1|e_1)$$

$$P(X_3|e_{1,}e_{2,}e_3) \propto P(e_3|X_3) \sum_{x_2} P(X_3|x_2) P(x_2|e_{1,}e_2)$$

…

Prediction: $P(X_{t+k} \mid e_{1:t})$, k>0

- $P(X_{t+1} \mid e_{1:t})$ part of the forward algorithm
- and from that on evidence does not count, and one can just calculate forward:

$$P(X_{t+2}|e_{1:t}) = \sum_{x_{t+1}} P(X_{t+2}|x_{t+1}, e_{1:t}) P(x_{t+1}|e_{1:t})$$

$$= \sum_{x_{t+1}} P(X_{t+2}|x_{t+1}) P(x_{t+1}|e_{1:t})$$

$$P(X_{t+3}|e_{1:t}) = \sum_{x_{t+2}} P(X_{t+3}|x_{t+2}, e_{1:t}) P(x_{t+2}|e_{1:t})$$

$$= \sum_{x_{t+2}} P(X_{t+3}|x_{t+2}) P(x_{t+2}|e_{1:t})$$

$$\ldots$$

Smoothing: $P(X_k \mid e_{1:t})$, k<t

- Obvious move: divide $e_{1:t}$ to $e_{1:k}$ and $e_{k+1:t}$.

$$\begin{aligned} P(X_k|e_{1:t}) &= P(X_k|e_{1:k}, e_{k+1:t}) \\ &\propto P(X_k|e_{1:k})\, P(e_{k+1:t}|X_k, e_{1:k}) \\ &= P(X_k|e_{1:k})\, \underline{P(e_{k+1:t}|X_k)} \end{aligned}$$

$$\begin{aligned} P(e_{k+1:t}|X_k) &= \sum_{x_{k+1}} P(x_{k+1}, e_{k+1:t}|X_k) \\ &= \sum_{x_{k+1}} P(x_{k+1}|X_k)\, P(e_{k+1:t}|x_{k+1}, X_k) \\ &= \sum_{x_{k+1}} P(x_{k+1}|X_k)\, P(e_{k+1}, e_{k+2:t}|x_{k+1}) \\ &= \sum_{x_{k+1}} P(x_{k+1}|X_k)\, P(e_{k+1}|x_{k+1})\, \underline{P(e_{k+2:t}|x_{k+1})} \end{aligned}$$

- and the first (last) step:

$$\begin{aligned} P(e_t|X_{t-1}) &= \sum_{x_t} P(x_t, e_t|X_{t-1}) = \sum_{x_t} P(e_t|x_t, X_{t-1})\, P(x_t|X_{t-1}) \\ &= \sum_{x_t} P(e_t|x_t)\, P(x_t|X_{t-1}) \end{aligned}$$

$$\begin{aligned}P(e_{k+1:t}|X_k) &= \sum_{x_{k+1}} P(x_{k+1}, e_{k+1:t}|X_k) \\ &= \sum_{x_{k+1}} P(x_{k+1}|X_k)\, P(e_{k+1:t}|x_{k+1}, X_k) \\ &= \sum_{x_{k+1}} P(x_{k+1}|X_k)\, P(e_{k+1}, e_{k+2:t}|x_{k+1}) \\ &= \sum_{x_{k+1}} P(x_{k+1}|X_k)\, P(e_{k+1}|x_{k+1})\, P(e_{k+2:t}|x_{k+1})\end{aligned}$$