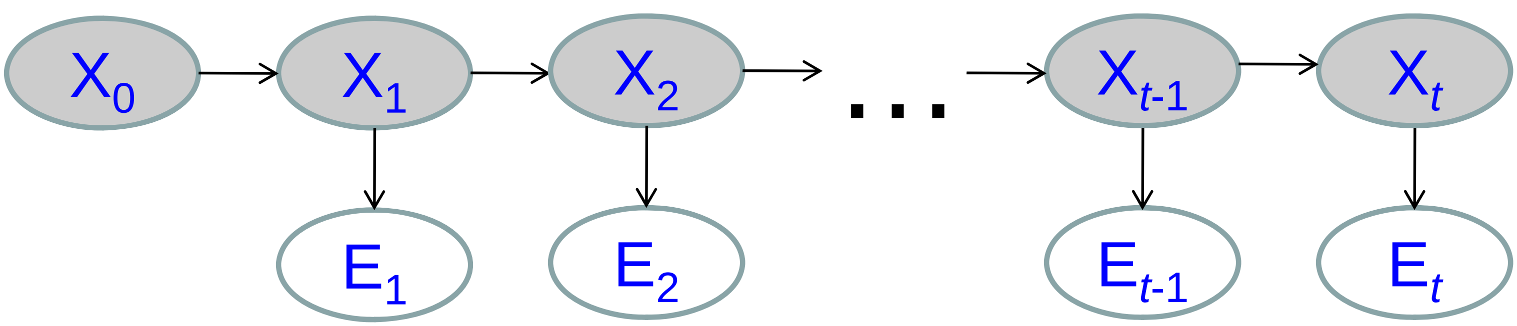
Hidden Markov Models

- **Markov assumption:**
  - The current state is conditionally independent of all the other past states given the state in the previous time step
  - The evidence at time *t* depends only on the state at time *t*
- **Transition model:**

$$P(X_t \mid \mathbf{X}_{0:t-1}) = P(X_t \mid X_{t-1})$$

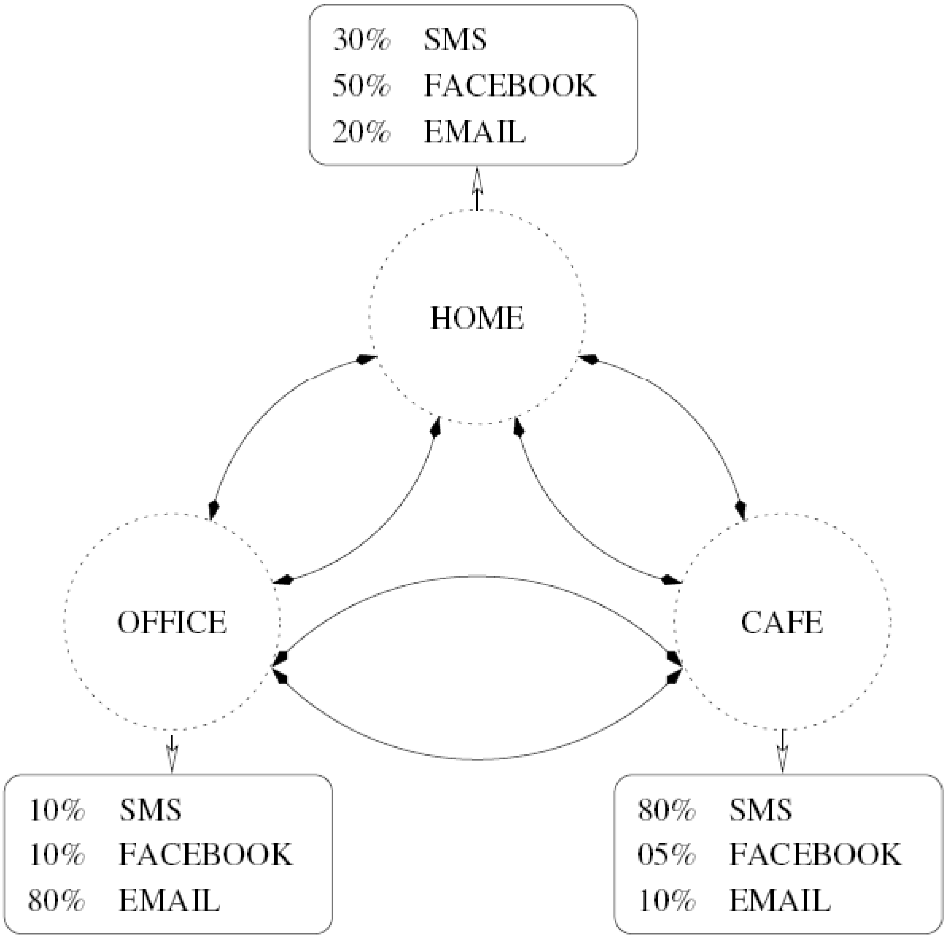
- **Observation model:**

$$P(E_t \mid \mathbf{X}_{0:t}, \mathbf{E}_{1:t-1}) = P(E_t \mid X_t)$$



An example HMM

- **States:** X = {home, office, cafe}
- **Observations:** E = {sms, facebook, email}



Transition Probabilities

	home	office	cafe
home	0.2	0.6	0.2
office	0.5	0.2	0.3
cafe	0.2	0.8	0.0

Emission Probabilities

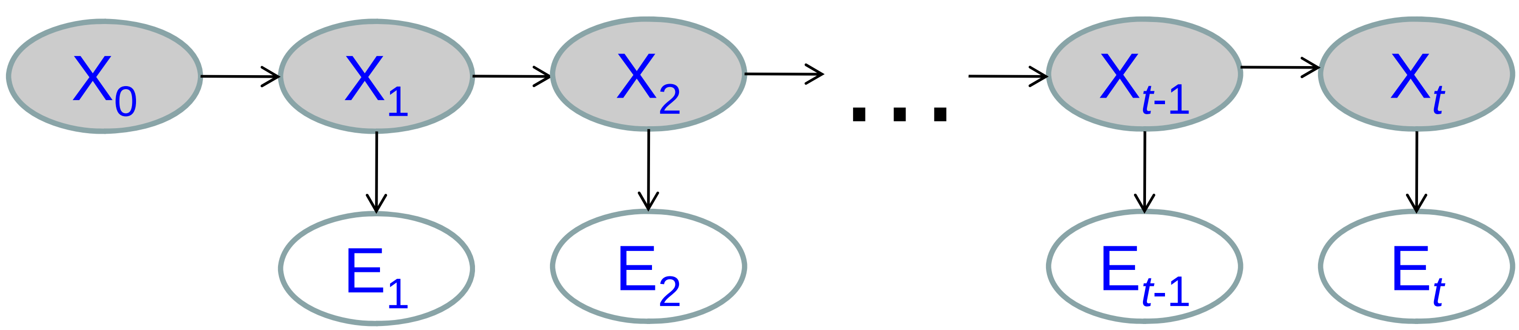
	sms	facebook	email
home	0.3	0.5	0.2
office	0.1	0.1	0.8
cafe	0.8	0.1	0.1

Slide credit: Andy White

The Joint Distribution

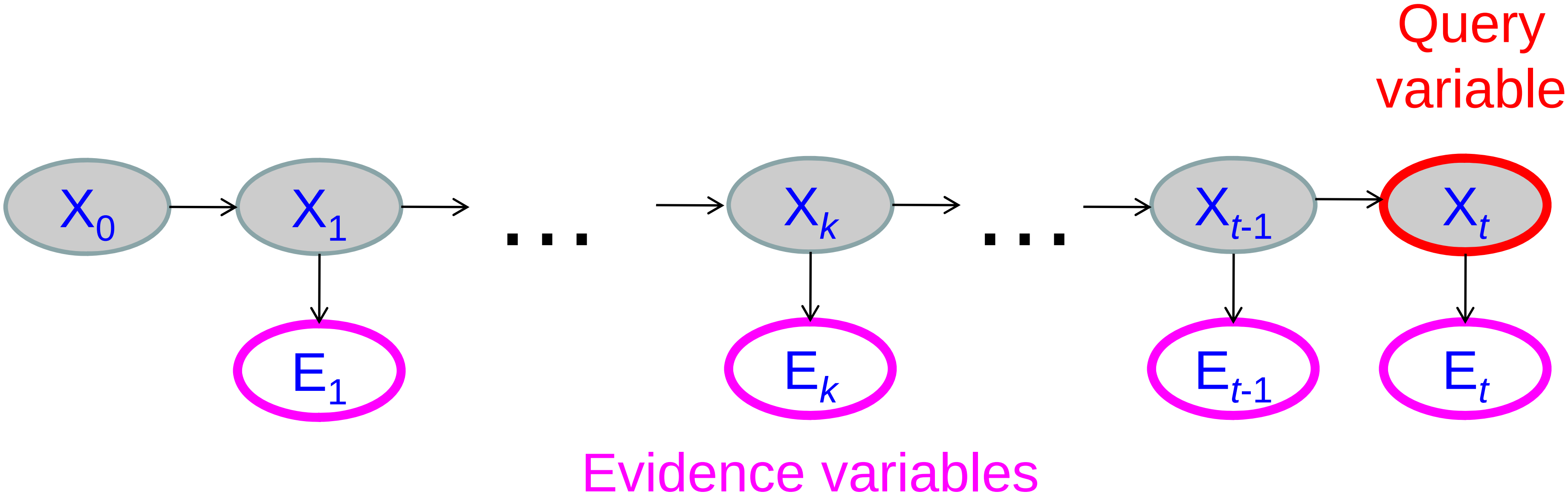
- Transition model: $P(X_t \mid X_{t-1})$
- Observation model: $P(E_t \mid X_t)$
- How do we compute the full joint $P(\mathbf{X}_{0:t}, \mathbf{E}_{1:t})$?

$$P(\boldsymbol{X}_{0:t}, \boldsymbol{E}_{1:t}) = P(X_0)\prod_{i=1}^{t} P(X_i | X_{i-1}) P(E_i | X_i)$$



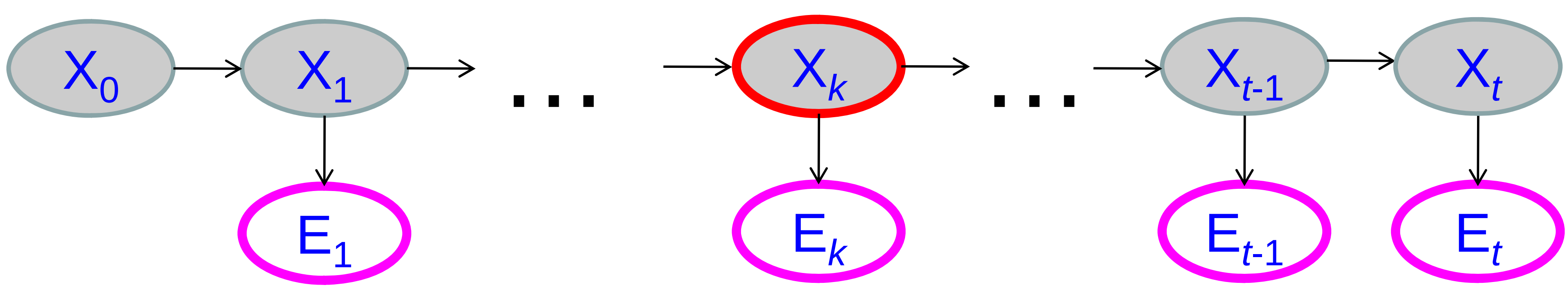
HMM inference tasks

- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{e}_{1:t}$?



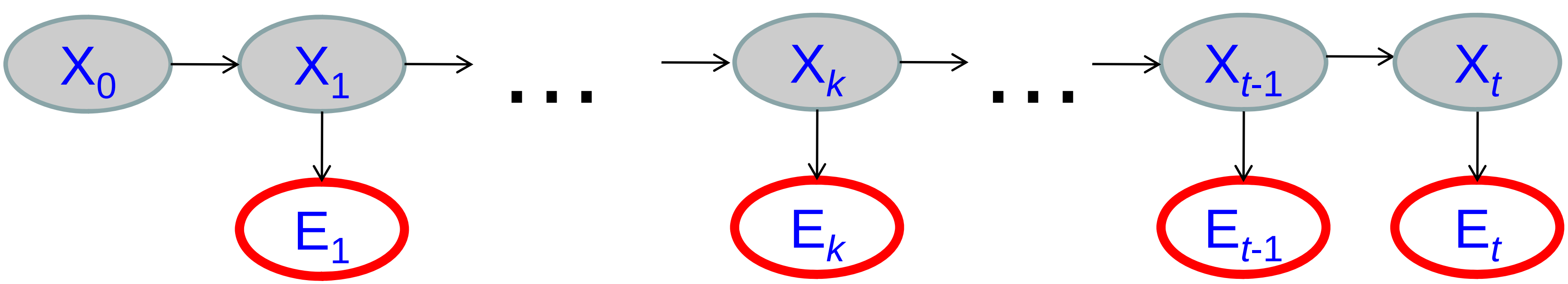
HMM inference tasks

- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{e}_{1:t}$?
- **Smoothing:** what is the distribution of some state X_k given the entire observation sequence $\mathbf{e}_{1:t}$?



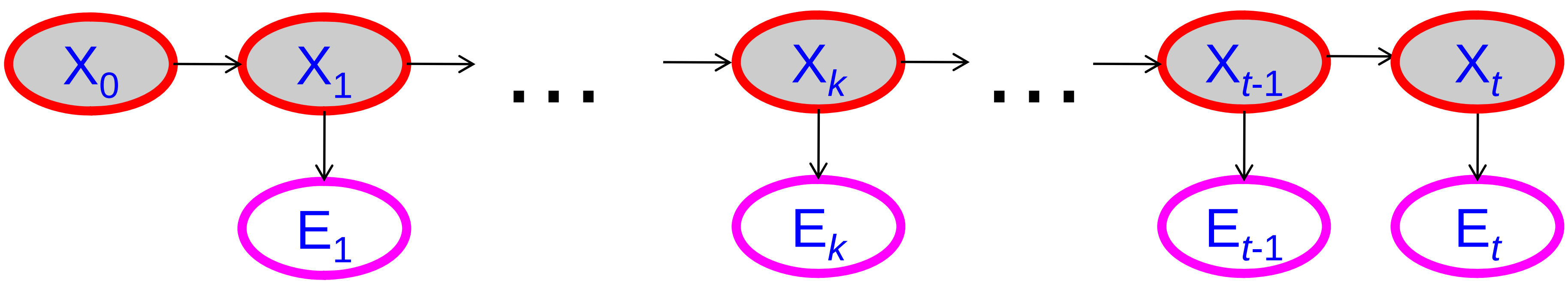
HMM inference tasks

- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{e}_{1:t}$?
- **Smoothing:** what is the distribution of some state X_k given the entire observation sequence $\mathbf{e}_{1:t}$?
- **Evaluation:** compute the probability of a given observation sequence $\mathbf{e}_{1:t}$



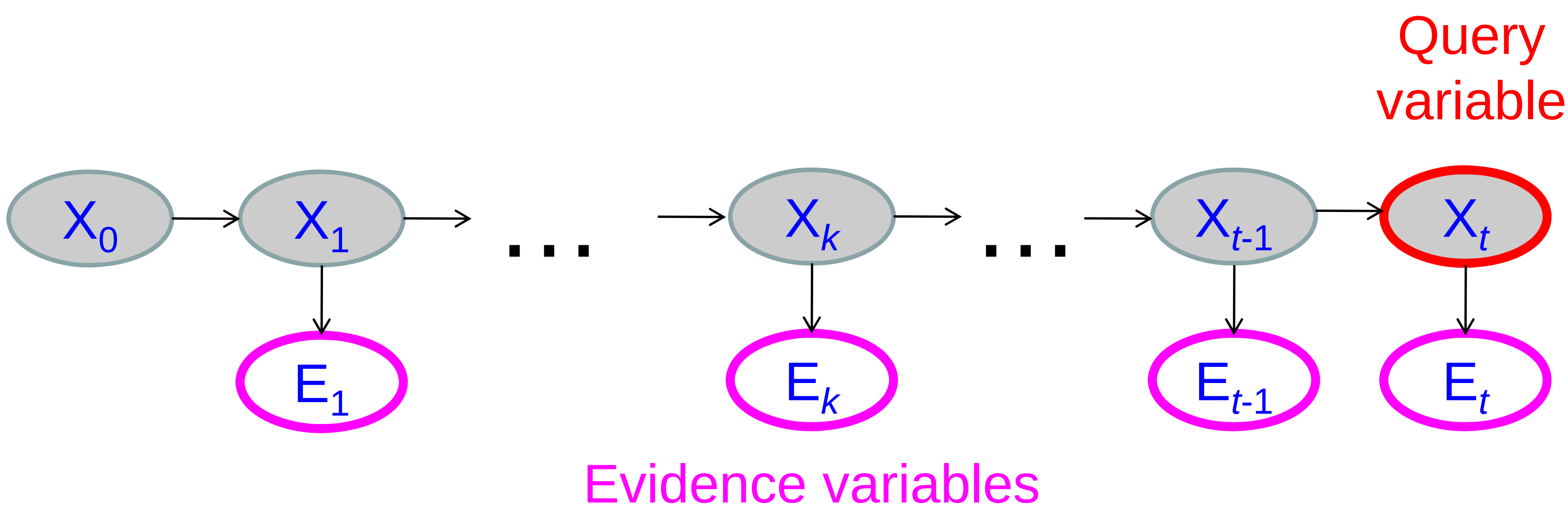
HMM inference tasks

- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{e}_{1:t}$
- **Smoothing:** what is the distribution of some state X_k given the entire observation sequence $\mathbf{e}_{1:t}$?
- **Evaluation:** compute the probability of a given observation sequence $\mathbf{e}_{1:t}$
- **Decoding:** what is the most likely state sequence $\mathbf{X}_{0:t}$ given the observation sequence $\mathbf{e}_{1:t}$?



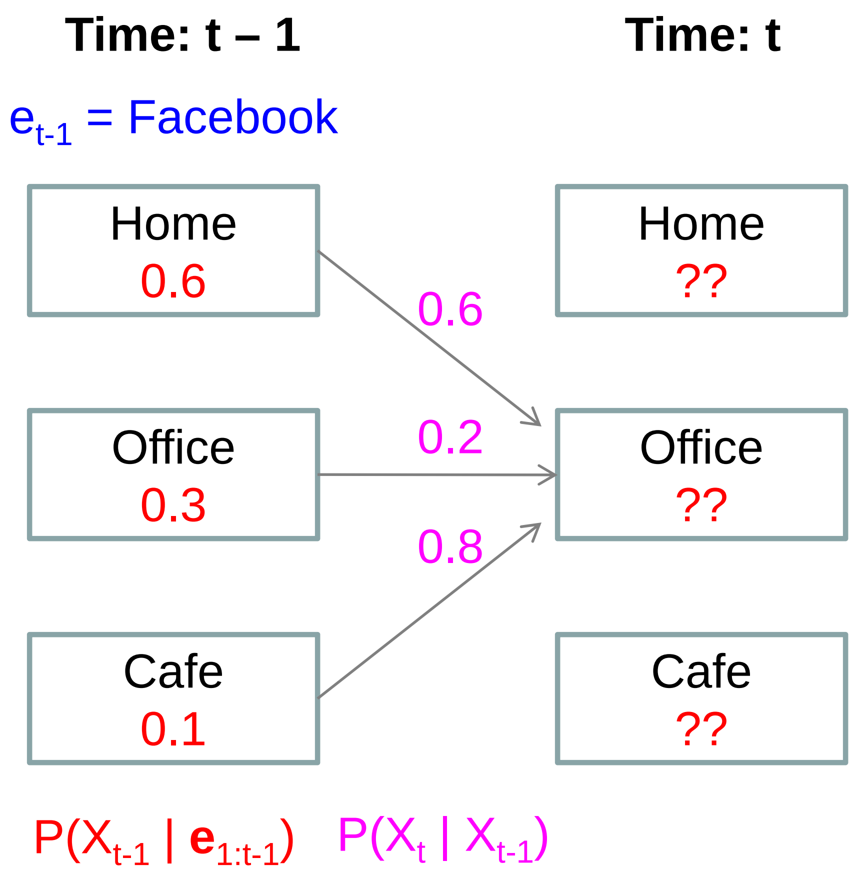
Filtering

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$



Filtering

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$



What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t-1})$?

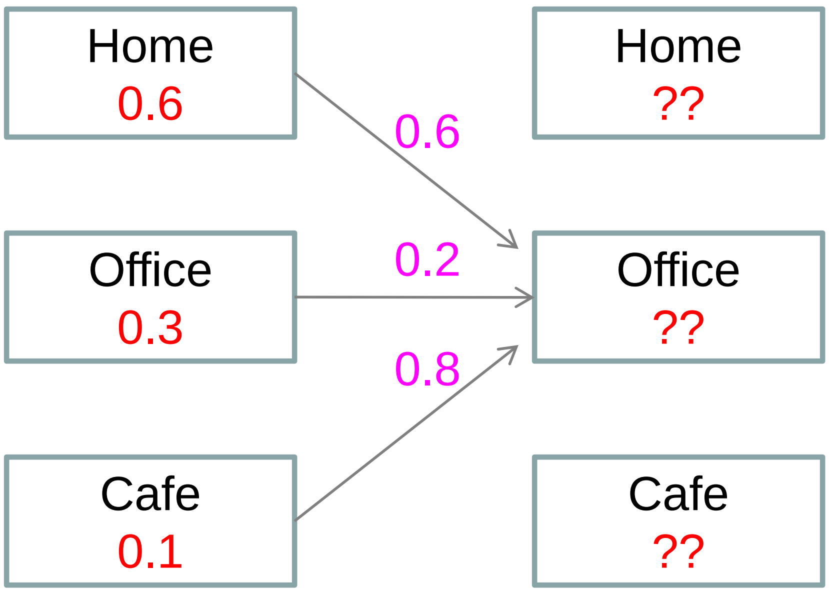
$0.6 * 0.6 + 0.2 * 0.3 + 0.8 * 0.1 = 0.5$

Filtering

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$

Time: t – 1 **Time: t**

e_{t-1} = Facebook



$P(X_{t-1} \mid \mathbf{e}_{1:t-1})$ $P(X_t \mid X_{t-1})$

What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t-1})$?

$0.6 * 0.6 + 0.2 * 0.3 + 0.8 * 0.1 = 0.5$

$$P(X_t \mid \boldsymbol{e}_{1:t-1}) = \sum_{x_{t-1}} P(X_t, x_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

$$= \sum_{x_{t-1}} P(X_t \mid x_{t-1}, \boldsymbol{e}_{1:t-1}) P(x_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

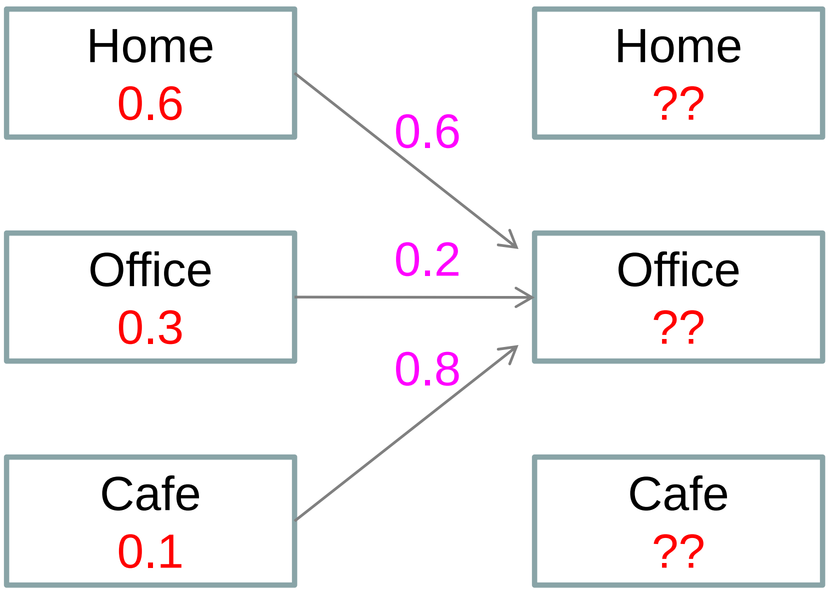
$$= \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

Filtering

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$

Time: t – 1 **Time: t**

e_{t-1} = Facebook



$P(X_{t-1} \mid \mathbf{e}_{1:t-1})$ $P(X_t \mid X_{t-1})$

What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t-1})$?

0.6 * 0.6 + 0.2 * 0.3 + 0.8 * 0.1 = 0.5

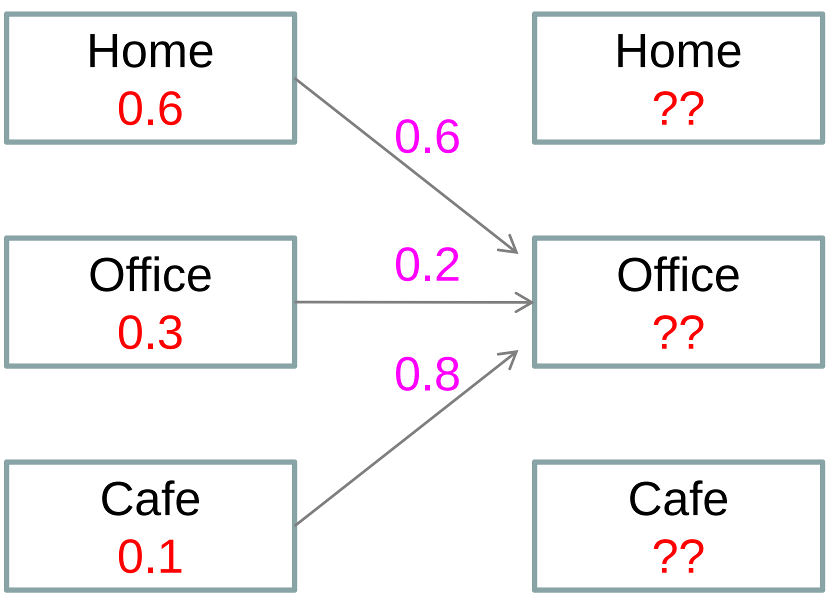
$$\boxed{P(X_t \mid \mathbf{e}_{1:t-1}) = \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid \mathbf{e}_{1:t-1})}$$

Filtering

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$

Time: t – 1 **Time: t**

e_{t-1} = Facebook e_t = Email



$P(X_{t-1} \mid \mathbf{e}_{1:t-1})$ $P(X_t \mid X_{t-1})$ $P(e_t \mid X_t) = 0.8$

What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t-1})$?

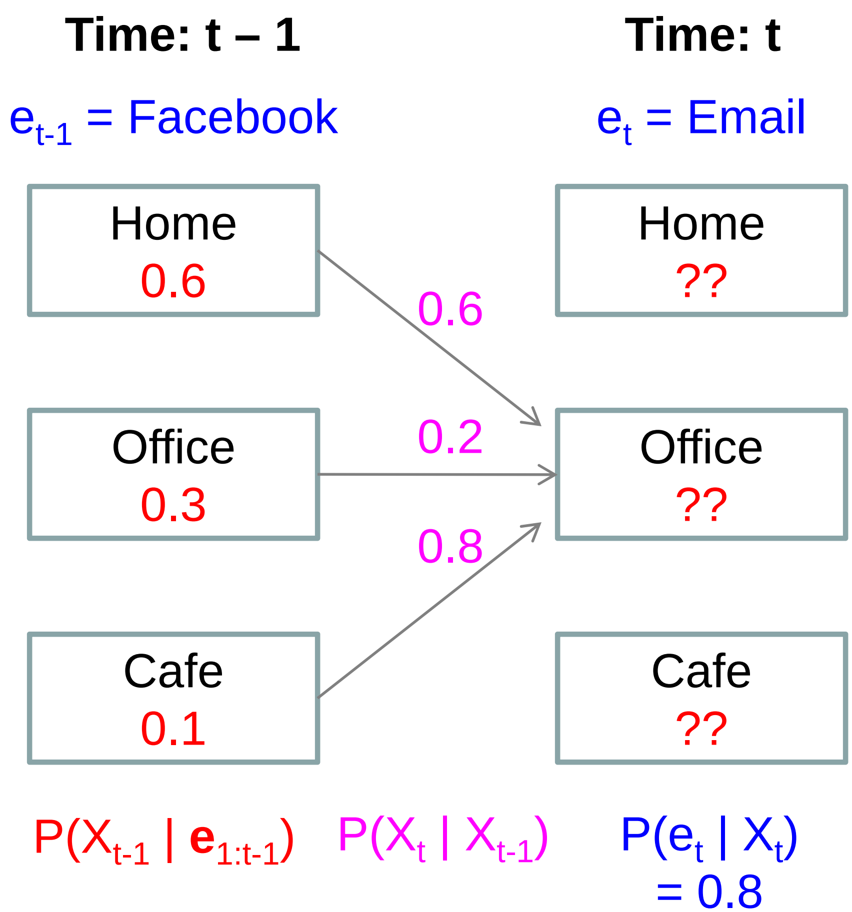
0.6 * 0.6 + 0.2 * 0.3 + 0.8 * 0.1 = 0.5

$$\boxed{P(X_t \mid \boldsymbol{e}_{1:t-1}) = \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid \boldsymbol{e}_{1:t-1})}$$

What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t})$?

Filtering

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$



What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t-1})$?

$0.6 * 0.6 + 0.2 * 0.3 + 0.8 * 0.1 = 0.5$

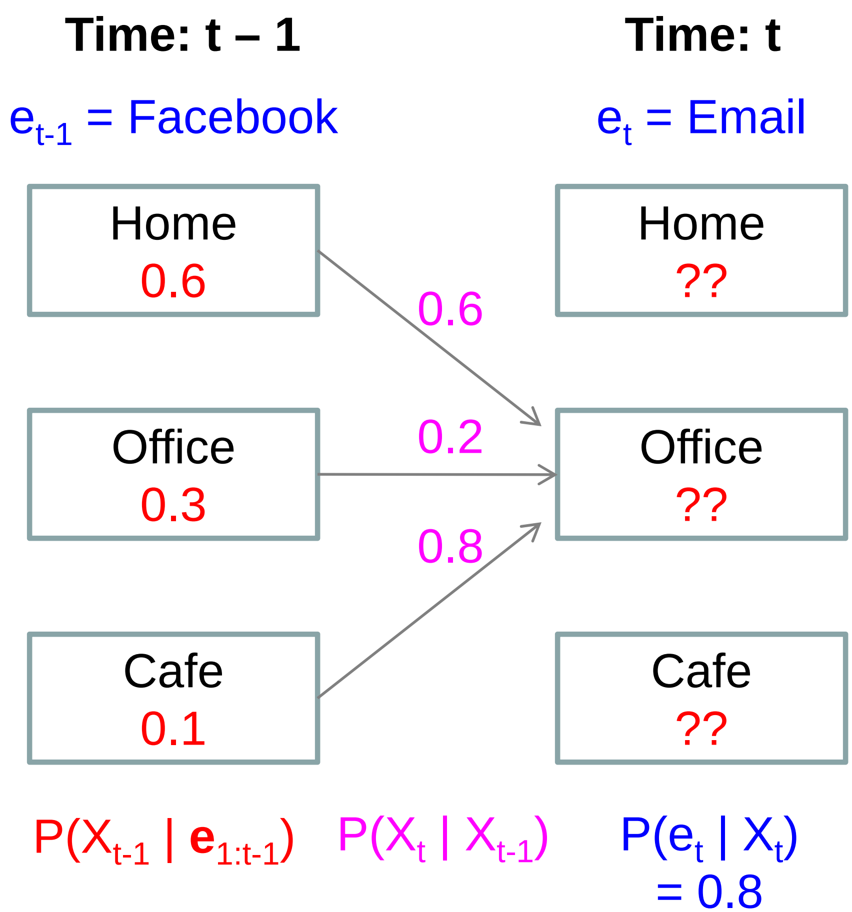
$$\boxed{P(X_t \mid \boldsymbol{e}_{1:t-1}) = \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid \boldsymbol{e}_{1:t-1})}$$

What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t})$?

$$P(X_t \mid e_t ; \boldsymbol{e}_{1:t-1})$$
$$= \frac{P(e_t \mid X_t ; \boldsymbol{e}_{1:t-1}) P(X_t \mid \boldsymbol{e}_{1:t-1})}{P(e_t \mid \boldsymbol{e}_{1:t-1})}$$
$$\propto P(e_t \mid X_t) P(X_t \mid \boldsymbol{e}_{1:t-1})$$

Filtering

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$



What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t-1})$?

$0.6 * 0.6 + 0.2 * 0.3 + 0.8 * 0.1 = 0.5$

$$\boxed{P(X_t \mid \mathbf{e}_{1:t-1}) = \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid \mathbf{e}_{1:t-1})}$$

What is $P(X_t = \text{Office} \mid \mathbf{e}_{1:t})$?

$$\boxed{P(X_t \mid \mathbf{e}_{1:t}) \propto P(e_t \mid X_t) P(X_t \mid \mathbf{e}_{1:t-1})}$$

$\propto 0.5 * 0.8 = 0.4$

Note: must also compute this value for Home and Cafe, and renormalize to sum to 1

Filtering: The Forward Algorithm

- Task: compute the probability distribution over the current state given all the evidence so far: $P(X_t \mid \mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(X_{t-1} \mid \mathbf{e}_{1:t-1})$
  - Base case: priors $P(X_0)$
- **Prediction:** propagate belief from X_{t-1} to X_t

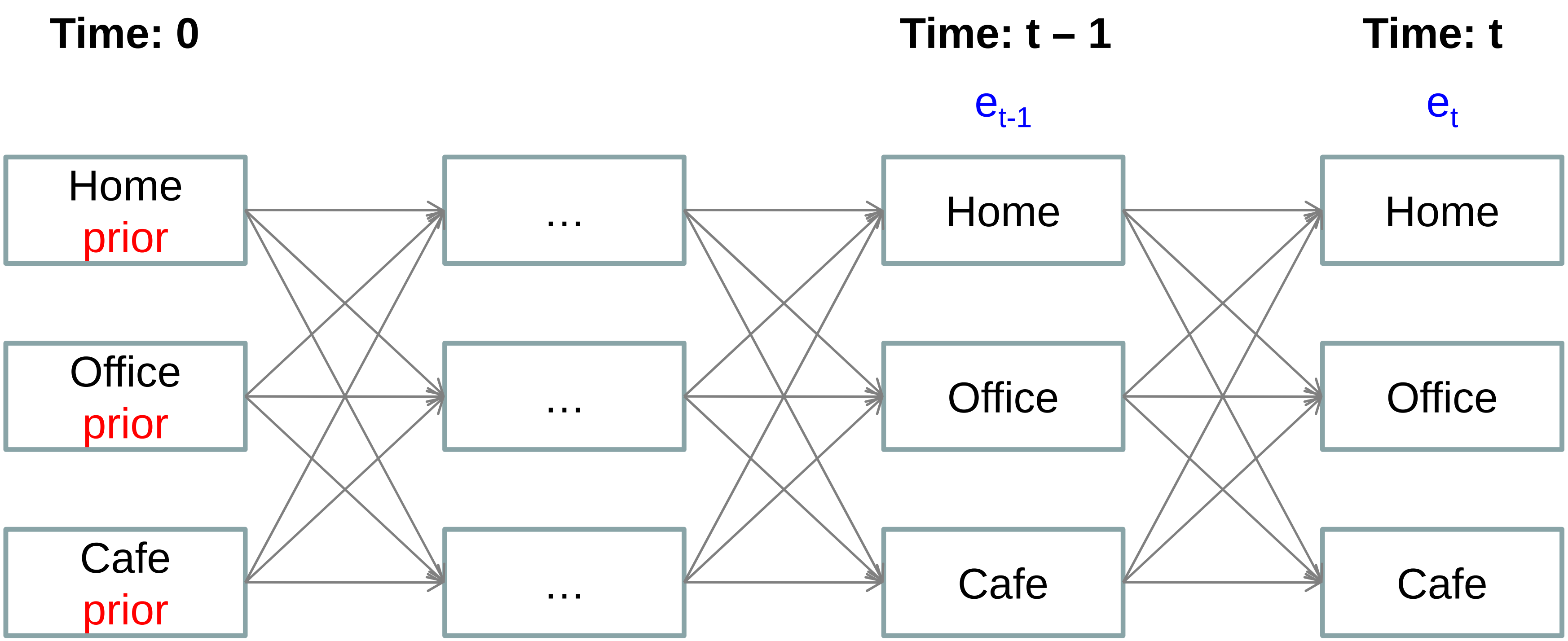
$$P(X_t \mid \boldsymbol{e}_{1:t-1}) = \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

- **Correction:** weight by evidence e_t

$$P(X_t \mid \boldsymbol{e}_{1:t}) = P(X_t \mid e_t ; \boldsymbol{e}_{1:t-1}) \propto P(e_t \mid X_t) P(X_t \mid \boldsymbol{e}_{1:t-1})$$

- Renormalize to have all $P(X_t = x \mid \mathbf{e}_{1:t})$ sum to 1

Filtering: The Forward Algorithm



Evaluation

- Compute the probability of the current sequence: $P(\mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(\mathbf{e}_{1:t-1})$

$$
\begin{aligned}
P(\boldsymbol{e}_{1:t}) &= P(\boldsymbol{e}_{1:t-1}, e_t) \\
&= P(\boldsymbol{e}_{1:t-1}) P(e_t \mid \boldsymbol{e}_{1:t-1}) \\
&= P(\boldsymbol{e}_{1:t-1}) \sum_{x_t} P(e_t, x_t \mid \boldsymbol{e}_{1:t-1}) \\
&= P(\boldsymbol{e}_{1:t-1}) \sum_{x_t} P(e_t \mid x_t, \boldsymbol{e}_{1:t-1}) P(x_t \mid \boldsymbol{e}_{1:t-1}) \\
&= P(\boldsymbol{e}_{1:t-1}) \sum_{x_t} P(e_t \mid x_t) P(x_t \mid \boldsymbol{e}_{1:t-1})
\end{aligned}
$$

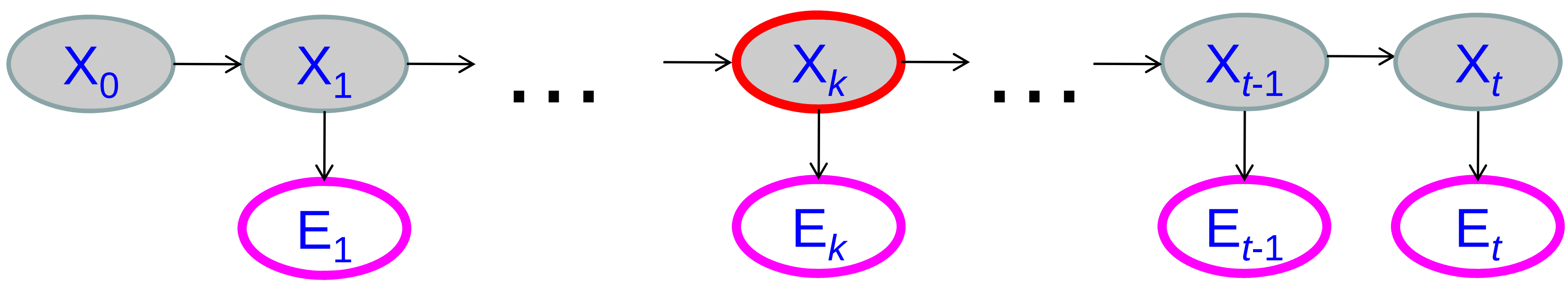
Evaluation

- Compute the probability of the current sequence: $P(\mathbf{e}_{1:t})$
- Recursive formulation: suppose we know $P(\mathbf{e}_{1:t-1})$

$$P(\boldsymbol{e}_{1:t}) = \underbrace{P(\boldsymbol{e}_{1:t-1})}_{\text{recursion}} \sum_{x_t} \underbrace{P(e_t \mid x_t) P(x_t \mid \boldsymbol{e}_{1:t-1})}_{\text{filtering}}$$

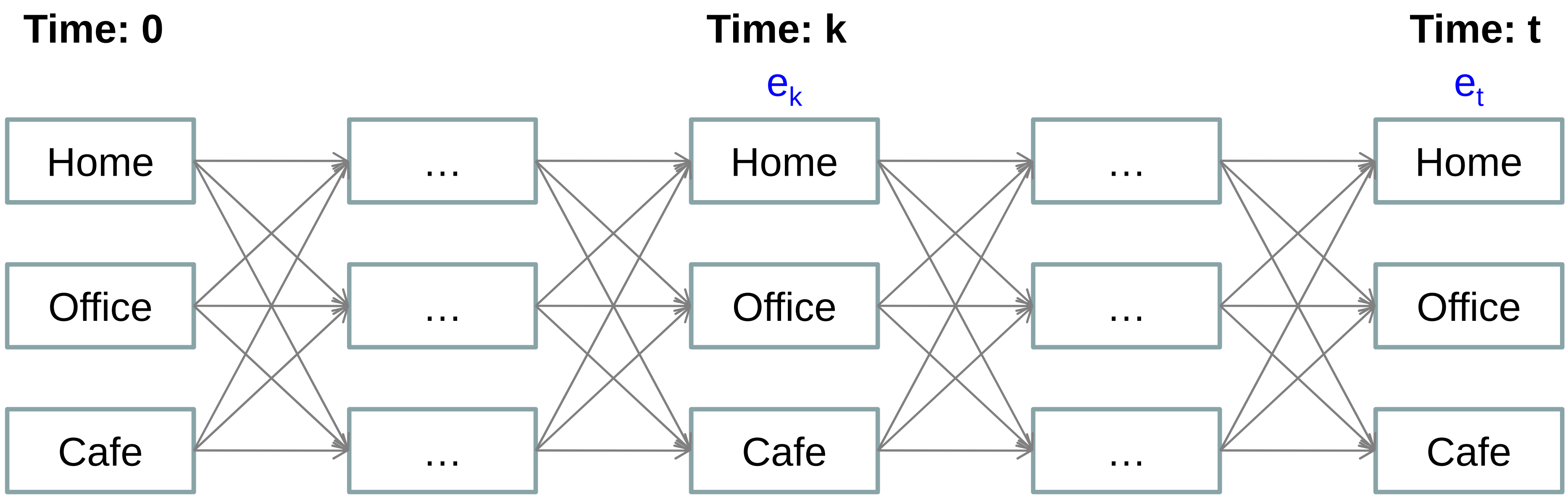
Smoothing

- What is the distribution of some state X_k given the entire observation sequence $\mathbf{e}_{1:t}$?



Smoothing

- What is the distribution of some state X_k given the entire observation sequence $\mathbf{e}_{1:t}$?
- Solution: the *forward-backward* algorithm



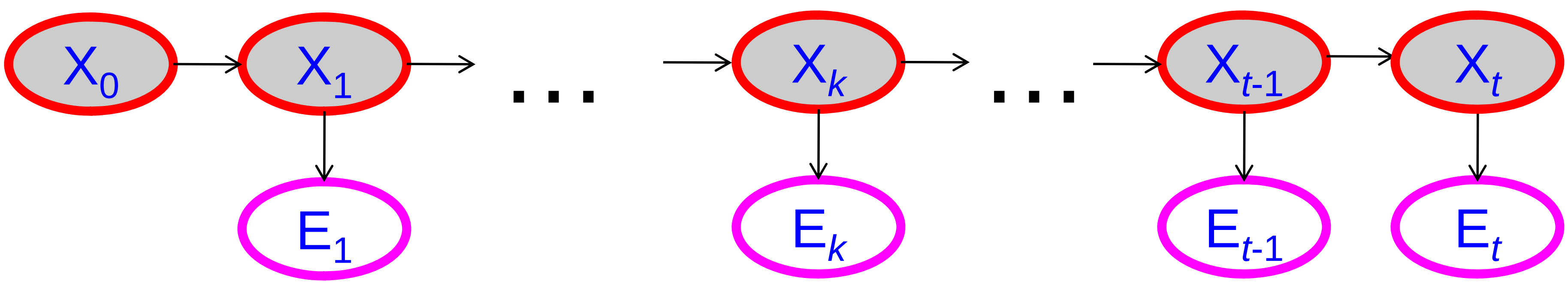
Forward message:
$P(X_k \mid \mathbf{e}_{1:k})$

Backward message:
$P(\mathbf{e}_{k+1:t} \mid X_k)$

Decoding: Viterbi Algorithm

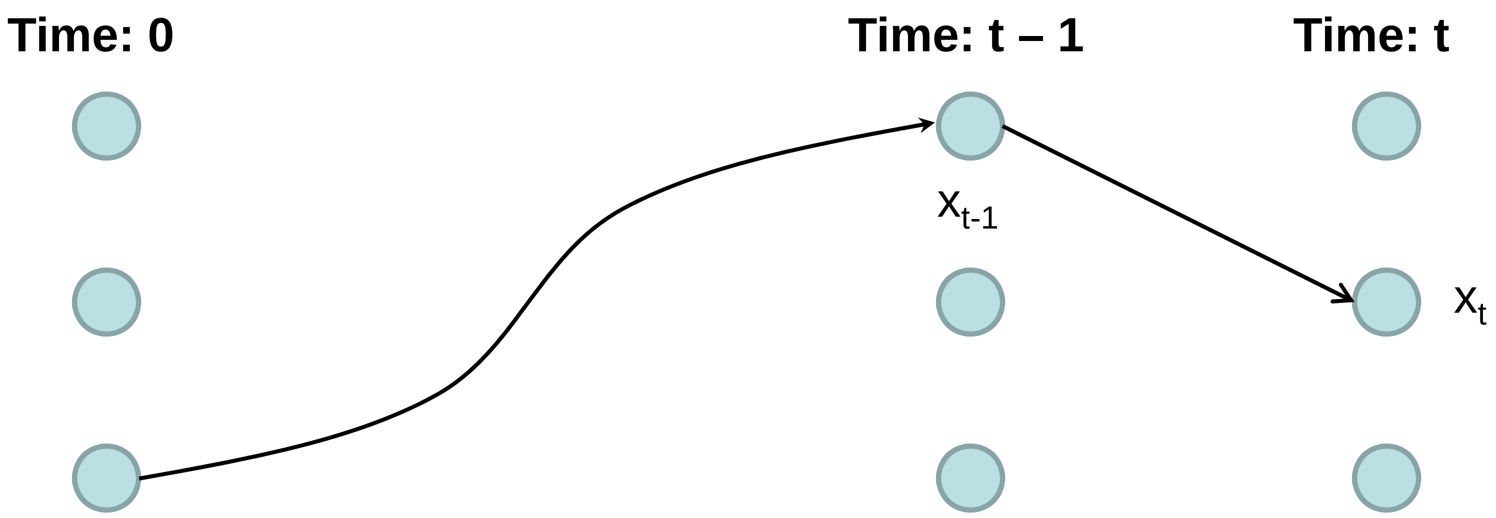
- Task: given observation sequence $\mathbf{e}_{1:t}$, compute most likely state sequence $\mathbf{x}_{0:t}$

$$\boldsymbol{x}^*_{0:t} = \arg\max_{\boldsymbol{x}_{0:t}} P(\boldsymbol{x}_{0:t} \mid \boldsymbol{e}_{1:t})$$



Decoding: Viterbi Algorithm

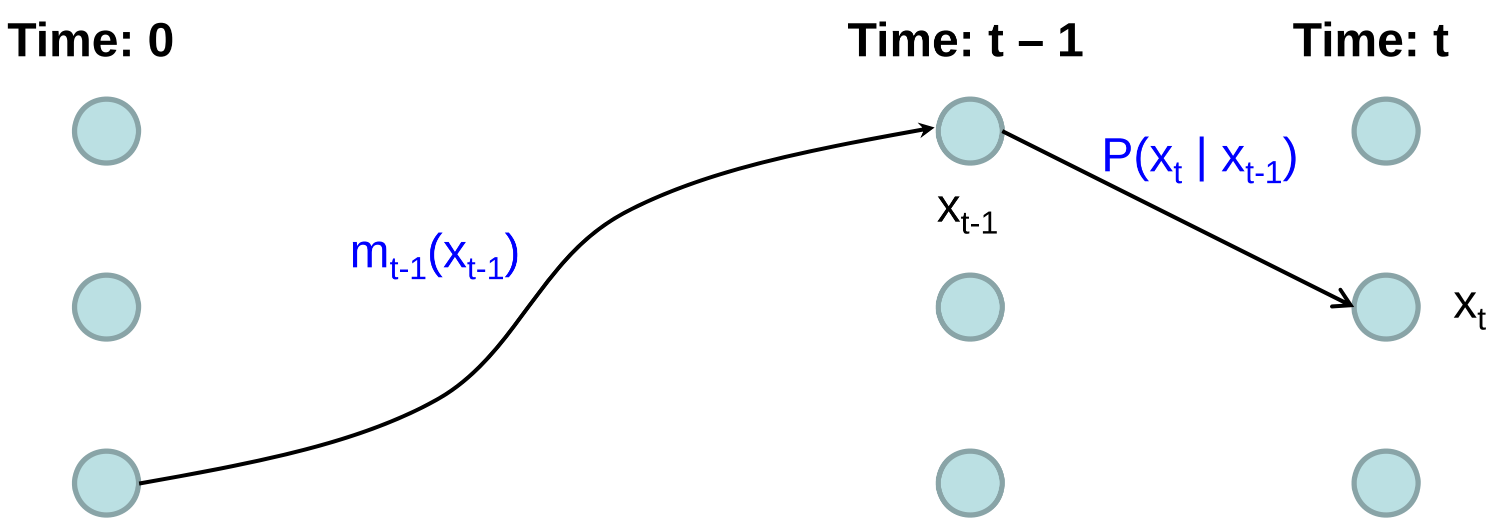
- Task: given observation sequence $\mathbf{e}_{1:t}$, compute most likely state sequence $\mathbf{x}_{0:t}$
- The most likely path that ends in a particular state x_t consists of the most likely path to some state x_{t-1} followed by the transition to x_t



Decoding: Viterbi Algorithm

- Let $m_t(x_t)$ denote the probability of the most likely path that ends in x_t:

$$
\begin{aligned}
m_t(x_t) &= \max_{\boldsymbol{x}_{0:t-1}} P(\boldsymbol{x}_{0:t-1}, x_t \mid \boldsymbol{e}_{1:t}) \\
&\propto \max_{\boldsymbol{x}_{0:t-1}} P(\boldsymbol{x}_{0:t-1}, x_t, \boldsymbol{e}_{1:t}) \\
&= \max_{x_{t-1}} \left[m_{t-1}(x_{t-1}) P(x_t \mid x_{t-1}) P(e_t \mid x_t) \right]
\end{aligned}
$$



Learning

- Given: a training sample of observation sequences
- Goal: compute model parameters
  - Transition probabilities $P(X_t \mid X_{t-1})$
  - Observation probabilities $P(E_t \mid X_t)$
- What if we had complete data, i.e., $\mathbf{e}_{1:t}$ and $\mathbf{x}_{0:t}$?
  - Then we could estimate all the parameters by relative frequencies

$$P(X_t = b \mid X_{t-1} = a) \approx \frac{\text{\# of times state b follows state a}}{\text{total \# of transitions from state a}}$$

$$P(E = e \mid X = a) \approx \frac{\text{\# of times e is emitted from state a}}{\text{total \# of emissions from state a}}$$

Learning

- Given: a training sample of observation sequences
- Goal: compute model parameters
  - Transition probabilities $P(X_t \mid X_{t\text{-}1})$
  - Observation probabilities $P(E_t \mid X_t)$
- What if we had complete data, i.e., $\mathbf{e}_{1:t}$ and $\mathbf{x}_{0:t}$?
  - Then we could estimate all the parameters by relative frequencies
- What if we knew the model parameters?
  - Then we could use inference to find the posterior distribution of the hidden states given the observations

Learning

- Given: a training sample of observation sequences
- Goal: compute model parameters
  - Transition probabilities $P(X_t \mid X_{t-1})$
  - Observation probabilities $P(E_t \mid X_t)$
- The **EM** (expectation-maximization) algorithm:

$$\theta^{(t+1)} = \arg\max_\theta \sum_{\boldsymbol{x}} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{e}, \theta^{(t)}) L(\boldsymbol{e}, \boldsymbol{X} = \boldsymbol{x} \mid \theta)$$

  - Starting with a random initialization of parameters:
    - **E-step:** find the posterior distribution of the hidden variables given observations and current parameter estimate
    - **M-step:** re-estimate parameter values given the expected values of the hidden variables