Conditional Models

Slides heavily borrowed from:
http://demo.clab.cs.cmu.edu/fa2015-11763/

Conditional Models

$$\mathcal{T} = (\langle \mathbf{x}_1, \mathbf{y}_1 \rangle, \langle \mathbf{x}_2, \mathbf{y}_2 \rangle, \ldots, \langle \mathbf{x}_n, \mathbf{y}_n \rangle)$$

Generative (joint) models(like HMMs) seek to maximize the following objective:

$$p(\mathcal{T}) = \prod_{\langle \mathbf{x}, \mathbf{y} \rangle \in \mathcal{T}} p(\mathbf{x}, \mathbf{y}; \boldsymbol{w})$$

Conditional models optimize
the following **conditional objective**

$$p(\mathcal{T}) = \prod_{\langle \mathbf{x}, \mathbf{y} \rangle \in \mathcal{T}} p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) \tilde{p}(\mathbf{x})$$

Why Conditional Models?

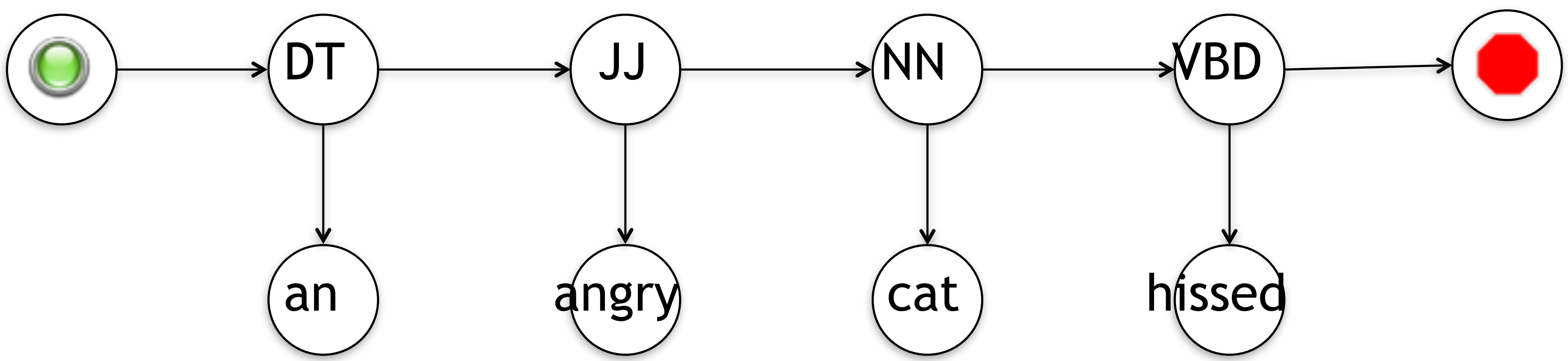
- Conditional models have the following property:

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sum_{\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}} p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) = 1$$

- Intuitively, we don't "waste" effort modeling the marginal distribution of **x**
- **HMMs** have restrictive expressive power because they try to model x with a simple/tractable model.

HMM recap

- Recall HMMs:



Posterior Marginals

- Marginal inference question for HMMs
  - Given **x**, what is the probability of being in a state *q* at time *i*?

$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \mathrm{START}) \times$$
$$p(x_{i+1}, \ldots, x_{|\mathbf{x}|} \mid y_i = q)$$

  - Given **x**, what is the probability of transitioning from state *q* to *r* at time *i*?

$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \mathrm{START}) \times$$
$$\eta(q \to r) \times \gamma(r \downarrow x_{i+1}) \times$$
$$p(x_{i+2}, \ldots, x_{|\mathbf{x}|} \mid y_{i+1} = r)$$

Posterior Marginals

- Marginal inference question for HMMs
  - Given **x**, what is the probability of being in a state *q* at time *i*?

$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \mathrm{START}) \times$$
$$p(x_{i+1}, \ldots, x_{|\mathbf{x}|} \mid y_i = q)$$

  - Given **x**, what is the probability of transitioning from state *q* to *r* at time *i*?

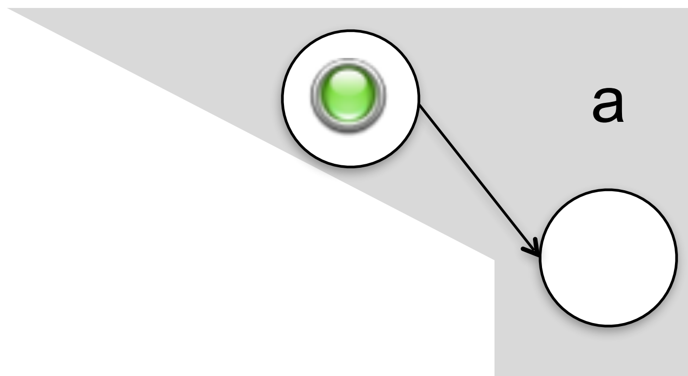
$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \mathrm{START}) \times$$
$$\eta(q \rightarrow r) \times \gamma(r \downarrow x_{i+1}) \times$$
$$p(x_{i+2}, \ldots, x_{|\mathbf{x}|} \mid y_{i+1} = r)$$

Forward Algorithm Recurrence

$$\alpha_0(\text{START}) = 1$$

$$\alpha_t(y) = \sum_{q \in \Omega} \eta(q \rightarrow y) \times \gamma(y \downarrow x_i) \times \alpha_{t-1}(q)$$

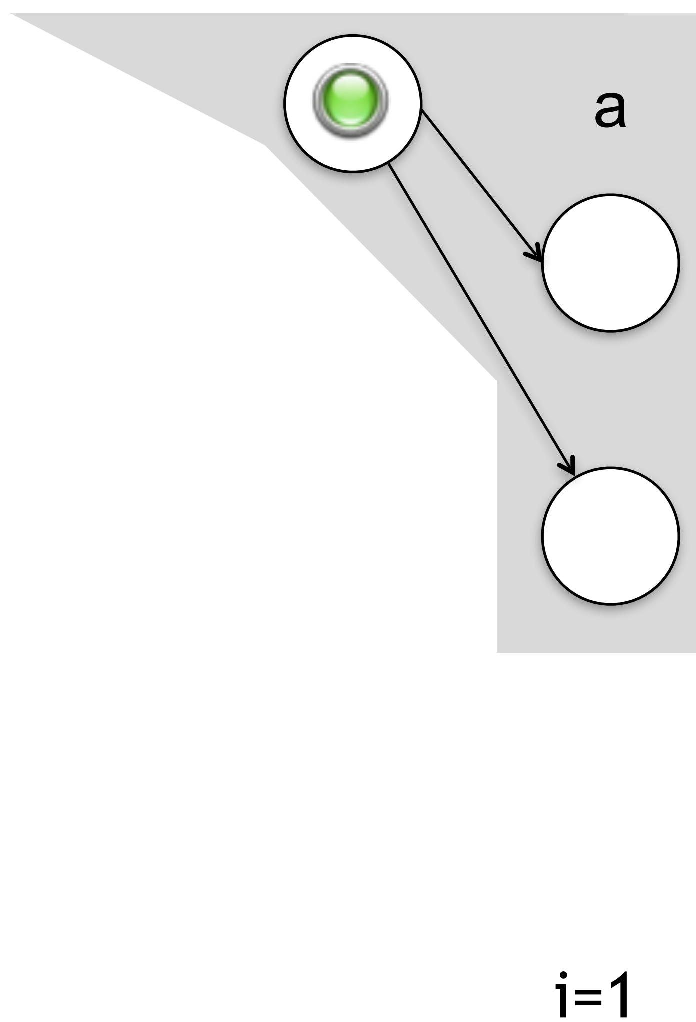
Forward Chart



i=1

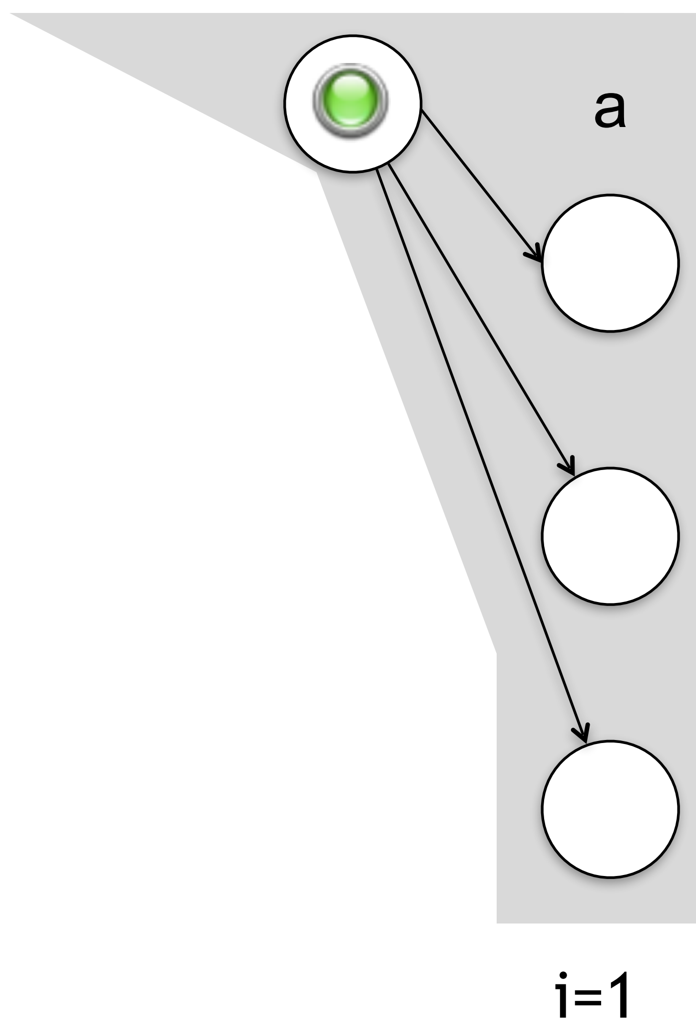
$$\alpha_t(q) = p(\mathrm{START}, x_1, \ldots, x_t, y_t = q)$$

Forward Chart



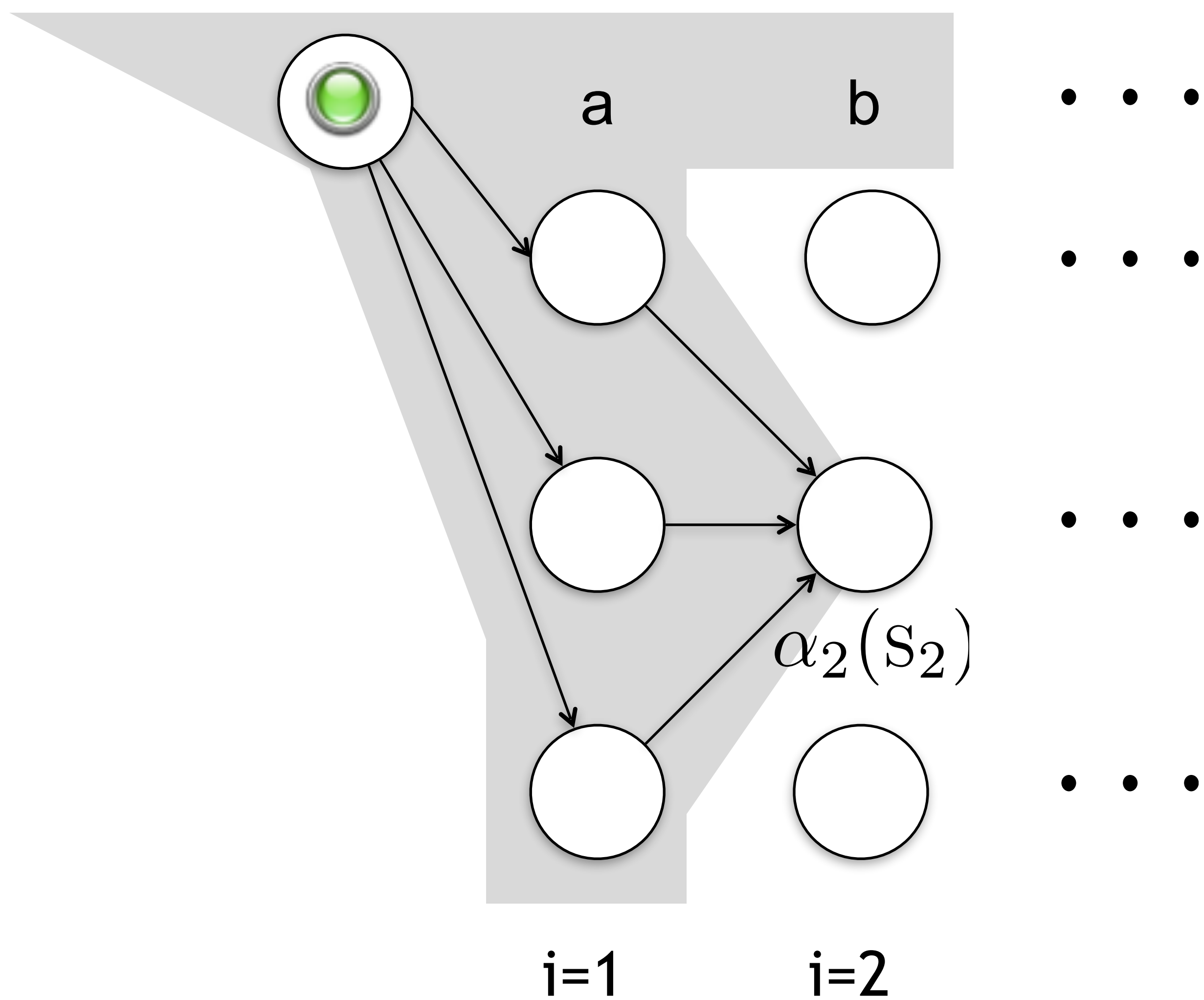
$$\alpha_t(q) = p(\mathrm{START}, x_1, \ldots, x_t, y_t = q)$$

Forward Chart



$$\alpha_t(q) = p(\text{START}, x_1, \ldots, x_t, y_t = q)$$

Forward Chart



$$\alpha_t(q) = p(\mathrm{START}, x_1, \ldots, x_t, y_t = q)$$

Posterior Marginals

- Marginal inference question for HMMs
  - Given **x**, what is the probability of being in a state *q* at time *i*?

$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \text{START}) \times$$

$$p(x_{i+1}, \ldots, x_{|\mathbf{x}|} \mid y_i = q)$$

  - Given **x**, what is the probability of transitioning from state *q* to *r* at time *i*?

$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \text{START}) \times$$

$$\eta(q \rightarrow r) \times \gamma(r \downarrow x_{i+1}) \times$$

$$p(x_{i+2}, \ldots, x_{|\mathbf{x}|} \mid y_{i+1} = r)$$

Backward Algorithm

- Start at the goal node(s) and work **backwards** through the hypergraph
- What is the probability in the goal node cell?
- What if there is more than one cell?
- What is the value of the axiom cell?

Backward Recurrence

$$\beta_{|\mathbf{x}|+1}(\text{STOP}) = 1$$

$$\beta_i(q) = \sum_{r \in \Omega} \beta_{i+1}(r) \times \gamma(r \downarrow x_{i+1}) \times \eta(q \rightarrow r)$$

Backward Chart

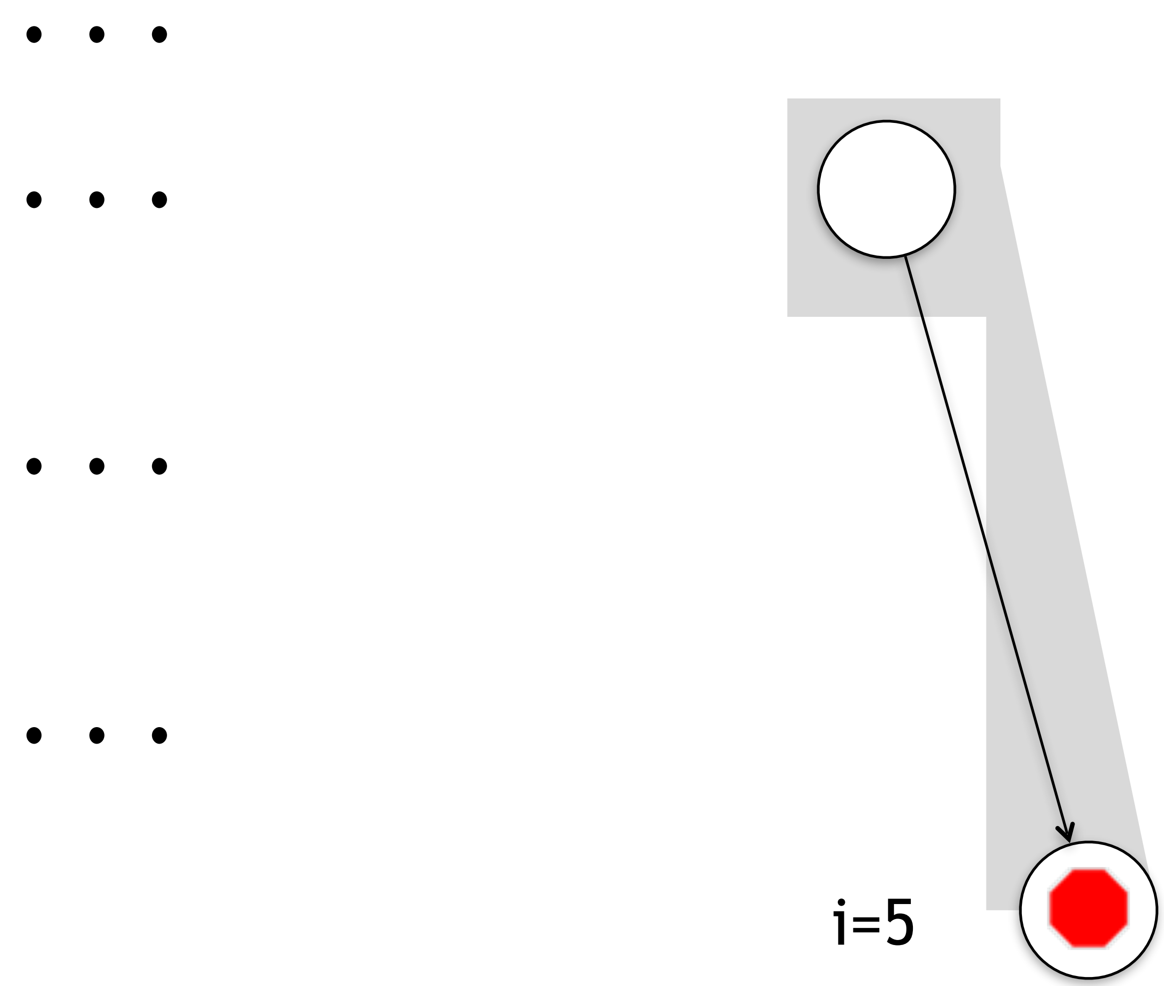
• • •

• • •

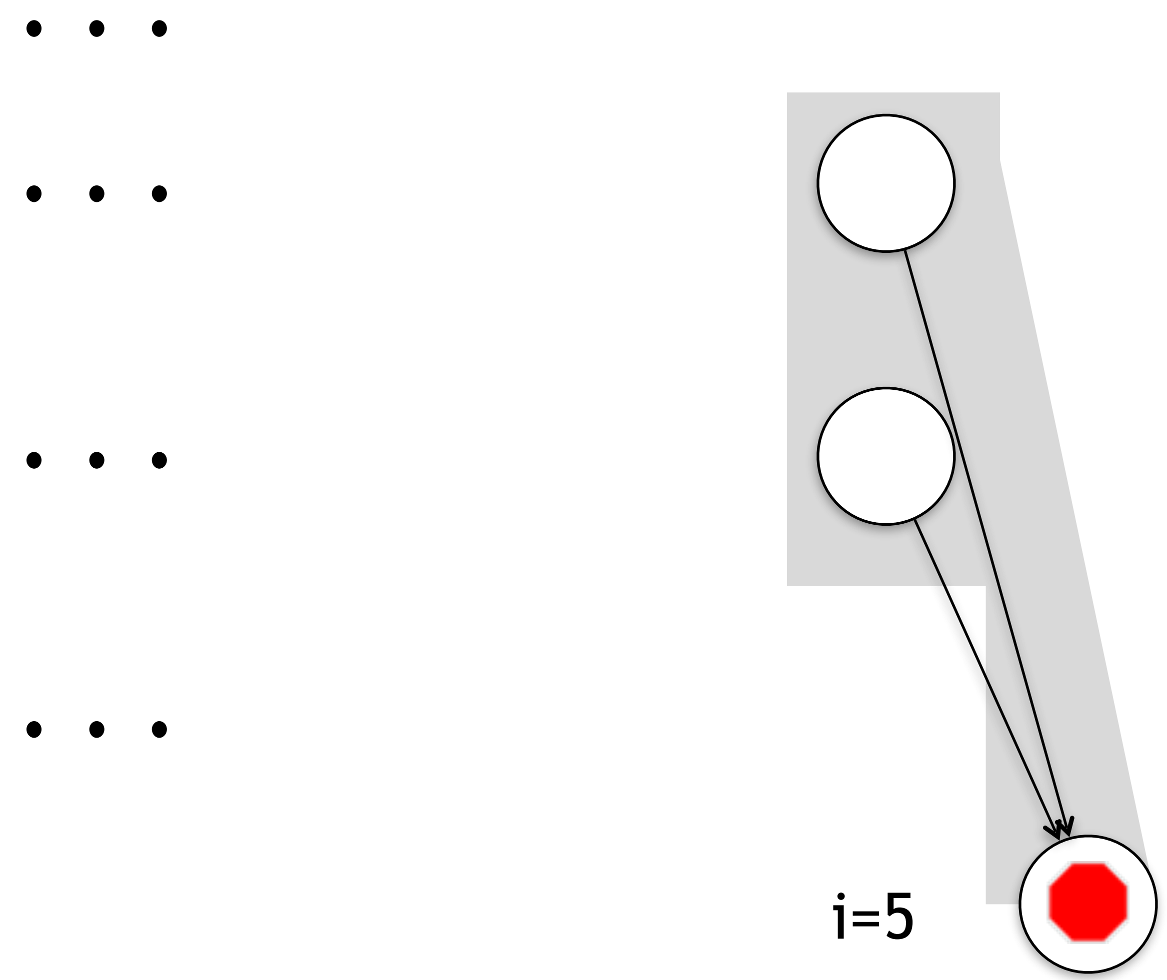
• • •

• • •

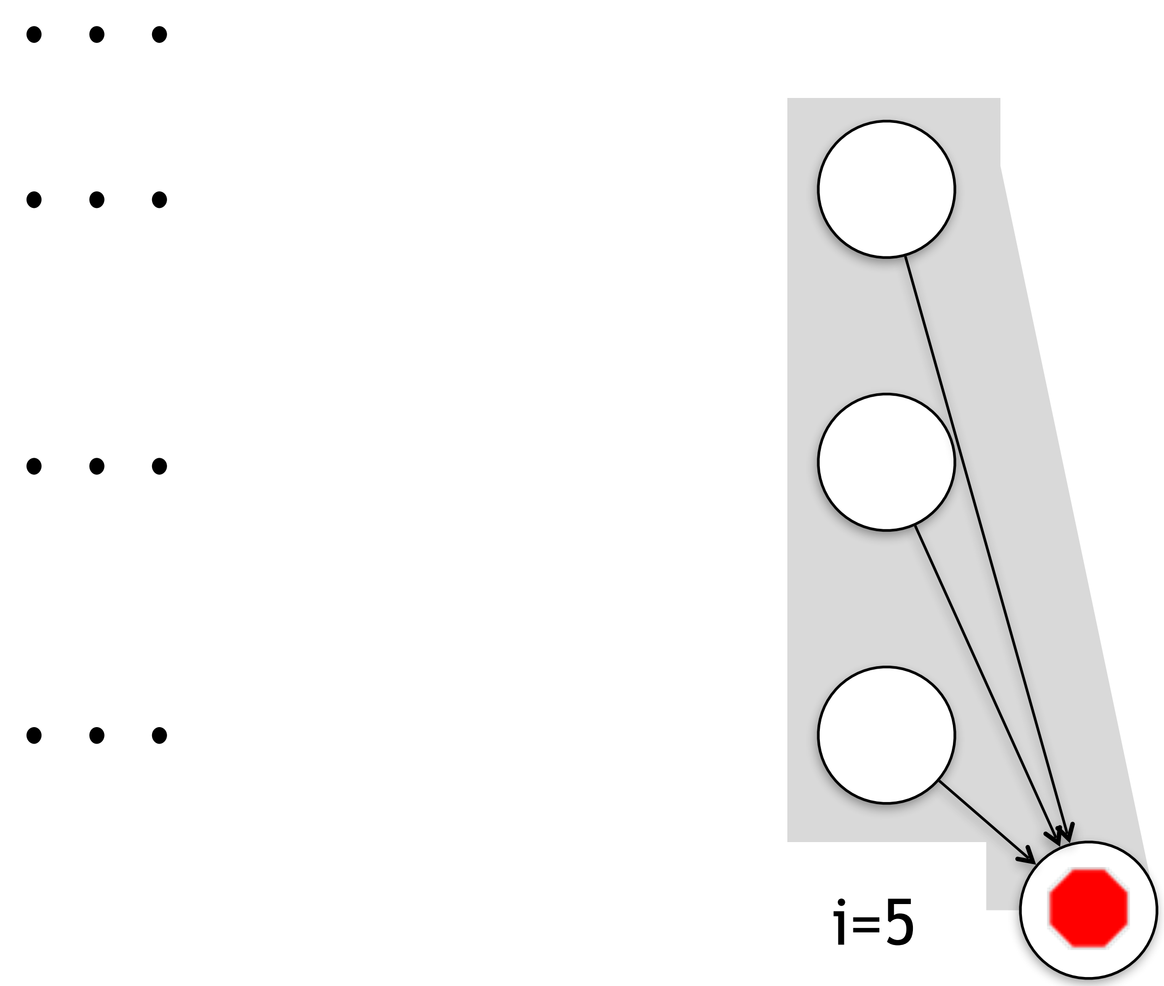
Backward Chart



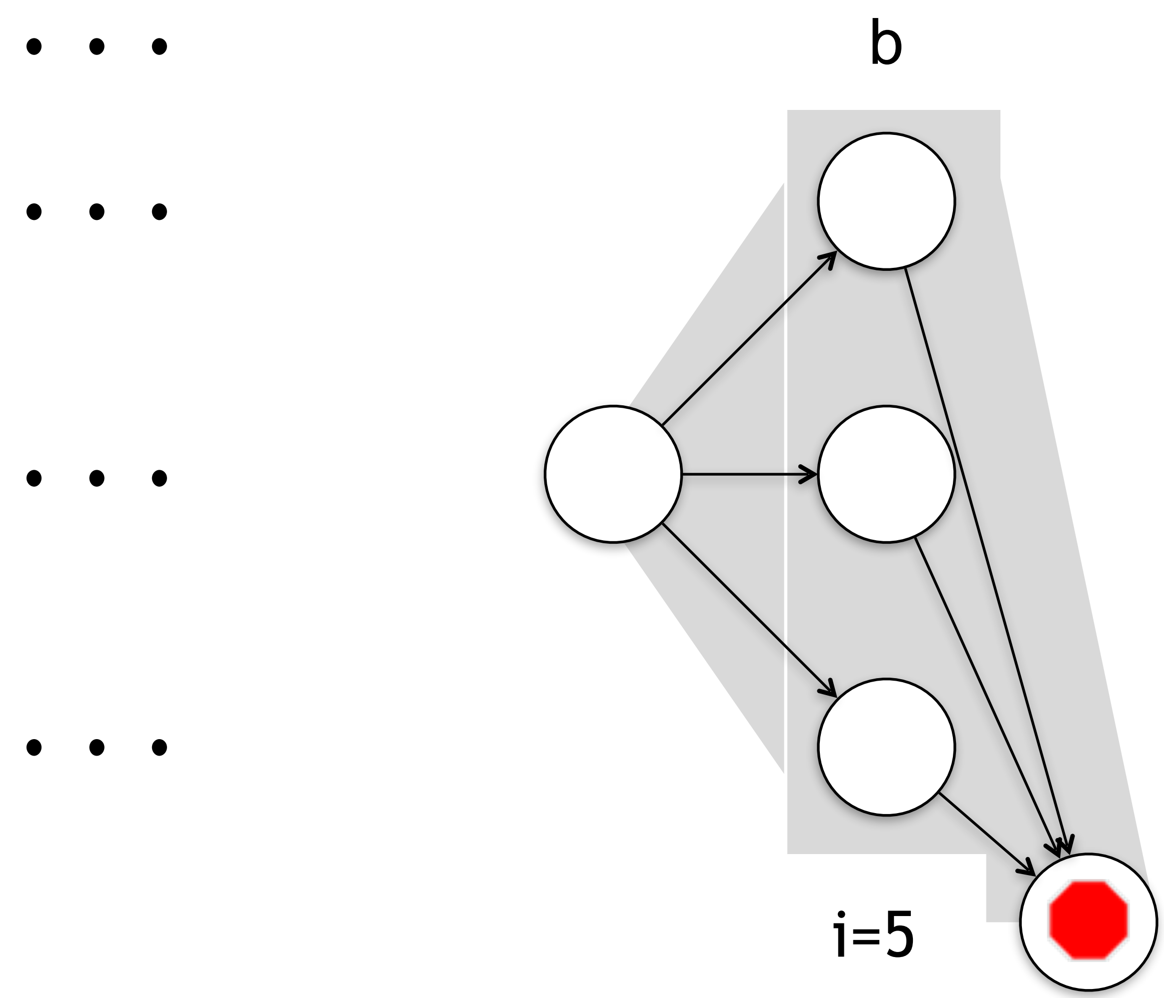
Backward Chart



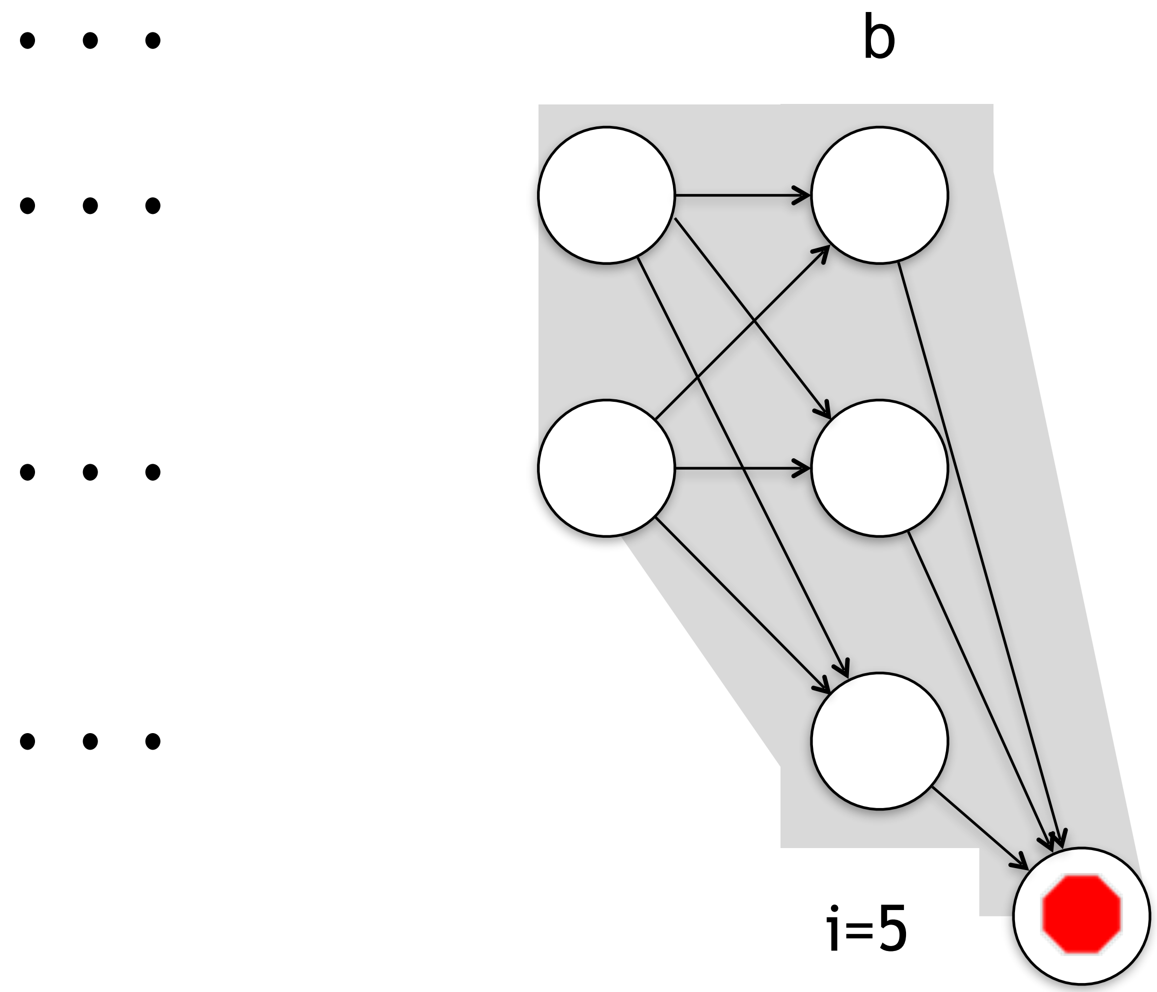
Backward Chart



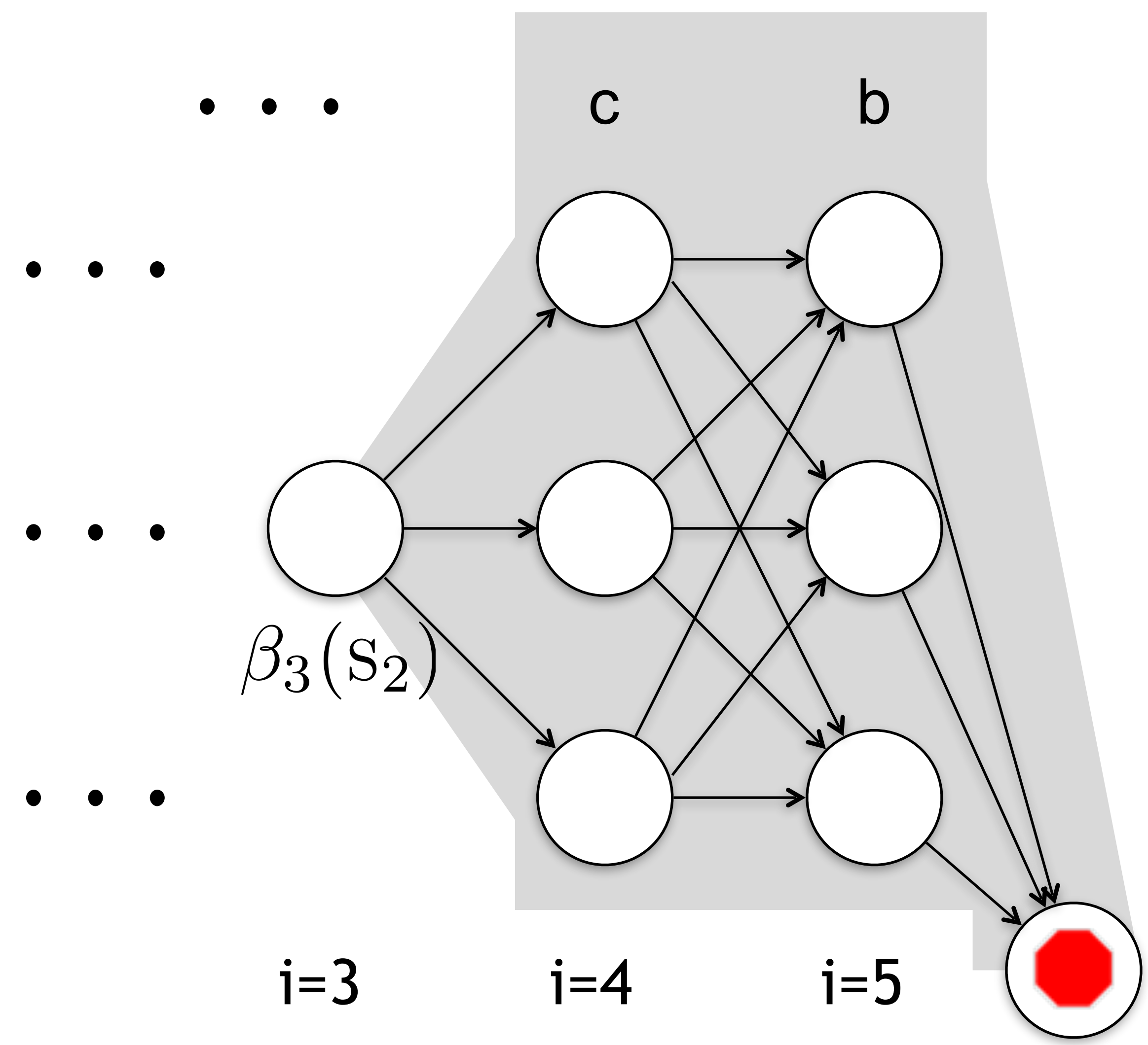
Backward Chart



Backward Chart



Backward Chart



$$\beta_t(q) = p(x_{t+1}, \ldots, x_{|\mathbf{x}|} \mid y_t = q)$$

Forward-Backward

- Compute forward chart

$$\alpha_t(q) = p(\text{START}, x_1, \ldots, x_t, y_t = q)$$

- Compute backward chart

$$\beta_t(q) = p(x_{t+1}, \ldots, x_{|\mathbf{x}|}, \text{STOP} \mid y_t = q)$$

What is $\alpha_t(q) \times \beta_t(q)$ **?**

Forward-Backward

- Compute forward chart

$$\alpha_t(q) = p(\text{START}, x_1, \ldots, x_t, y_t = q)$$

- Compute backward chart

$$\beta_t(q) = p(x_{t+1}, \ldots, x_{|\mathbf{x}|}, \text{STOP} \mid y_t = q)$$

What is $\alpha_t(q) \times \beta_t(q)$ **?**

$$p(\mathbf{x}, y_t = q) = \alpha_t(q) \times \beta_t(q)$$

Edge Marginals

- What is the probability that **x** was generated and q -> r happened at time t?

$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \text{START}) \times$$
$$\eta(q \to r) \times \gamma(r \downarrow x_{i+1}) \times$$
$$p(x_{i+2}, \ldots, x_{|\mathbf{x}|} \mid y_{i+1} = r)$$

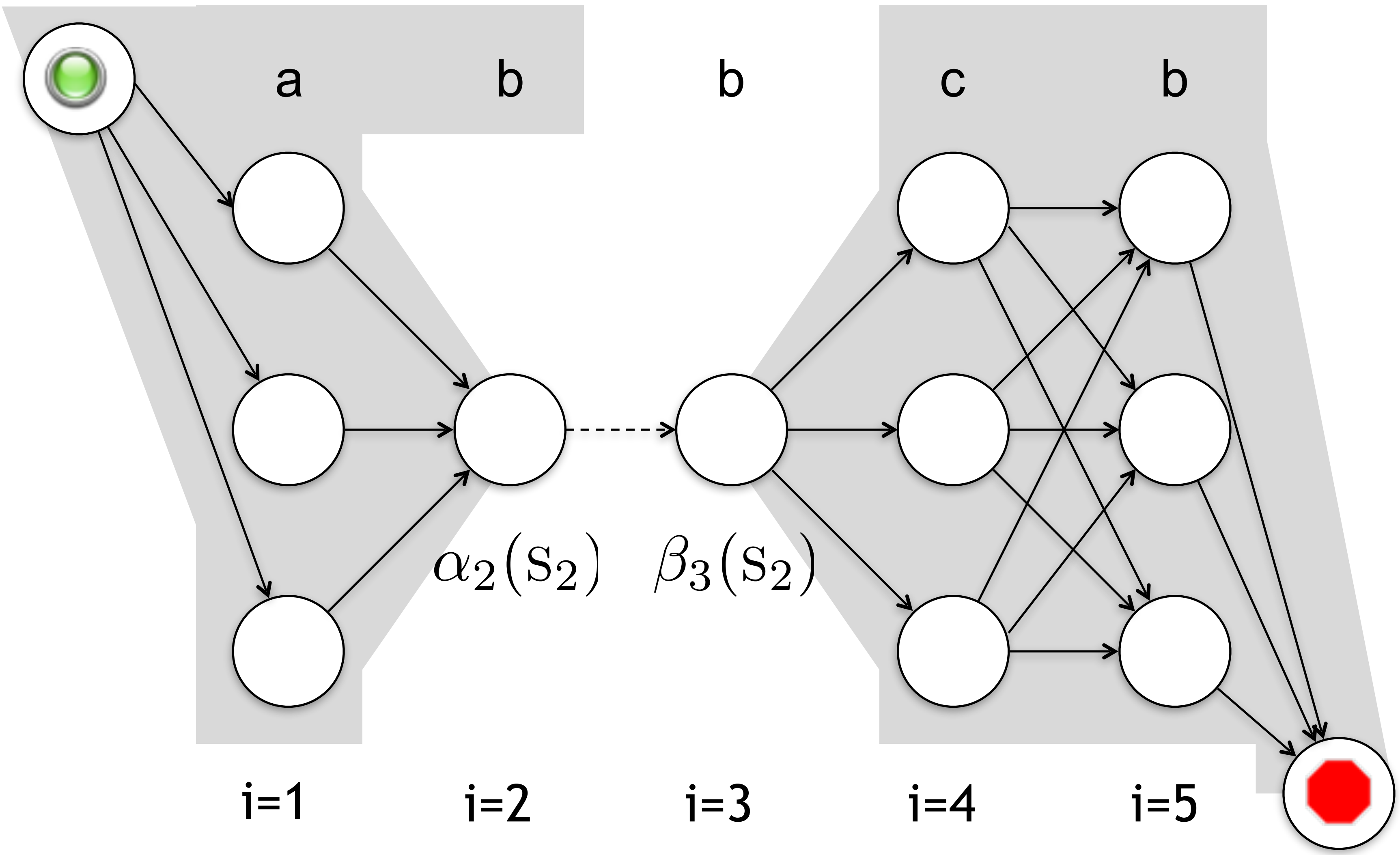
Edge Marginals

- What is the probability that **x** was generated and q -> r happened at time *t*?

$$p(x_1, \ldots, x_i, y_i = q \mid y_0 = \text{START}) \times$$
$$\eta(q \rightarrow r) \times \gamma(r \downarrow x_{i+1}) \times$$
$$p(x_{i+2}, \ldots, x_{|\mathbf{x}|} \mid y_{i+1} = r)$$

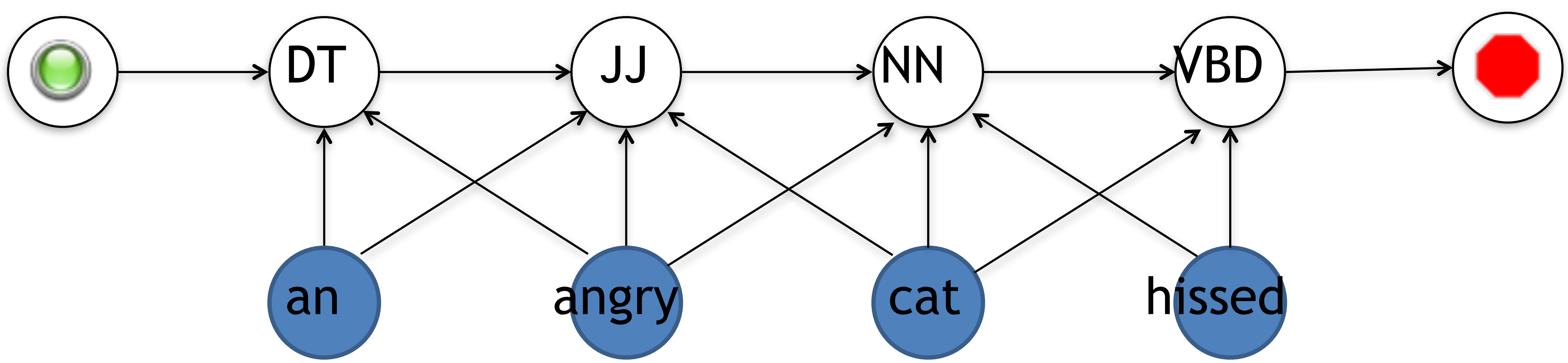
$$\alpha_t(q) \times$$
$$\eta(q \rightarrow r) \times \gamma(r \downarrow x_{t+1}) \times$$
$$\beta_{t+1}(r)$$

Forward-Backward



MEMMs

- Back to conditional modelling:



- **Limitation**: you cannot condition on the future, the probability p(**y** | **x**) still factors into conditionally independent steps

MEMM Structure

- MEMMs parameterize each local classification decision with a “conditional maximum entropy model” – more commonly known as a ***multiclass logistic regression classifie***r

$$p(y_i \mid \mathbf{x}, i, y_{i-1}; \boldsymbol{w}) = \frac{\exp \boldsymbol{w}^\top \boldsymbol{f}(y_i, \mathbf{x}, i, y_{i-1})}{\sum_{y' \in \Lambda} \exp \boldsymbol{w}^\top \boldsymbol{f}(y', \mathbf{x}, i, y_{i-1})}$$

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) = \prod_{i=1}^{|\mathbf{x}|} p(y_i \mid \mathbf{x}, i, y_{i-1}; \boldsymbol{w})$$

Learning MEMM Params

- The training objective is the conditional likelihood of all of the local classification decisions

$$\mathcal{L} = \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in \mathcal{T}} \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{w}^\top \boldsymbol{f}(y_i, \mathbf{x}, i, y_{i-1}) - \log Z(\mathbf{x}, i, y_{i-1}; \boldsymbol{w})$$

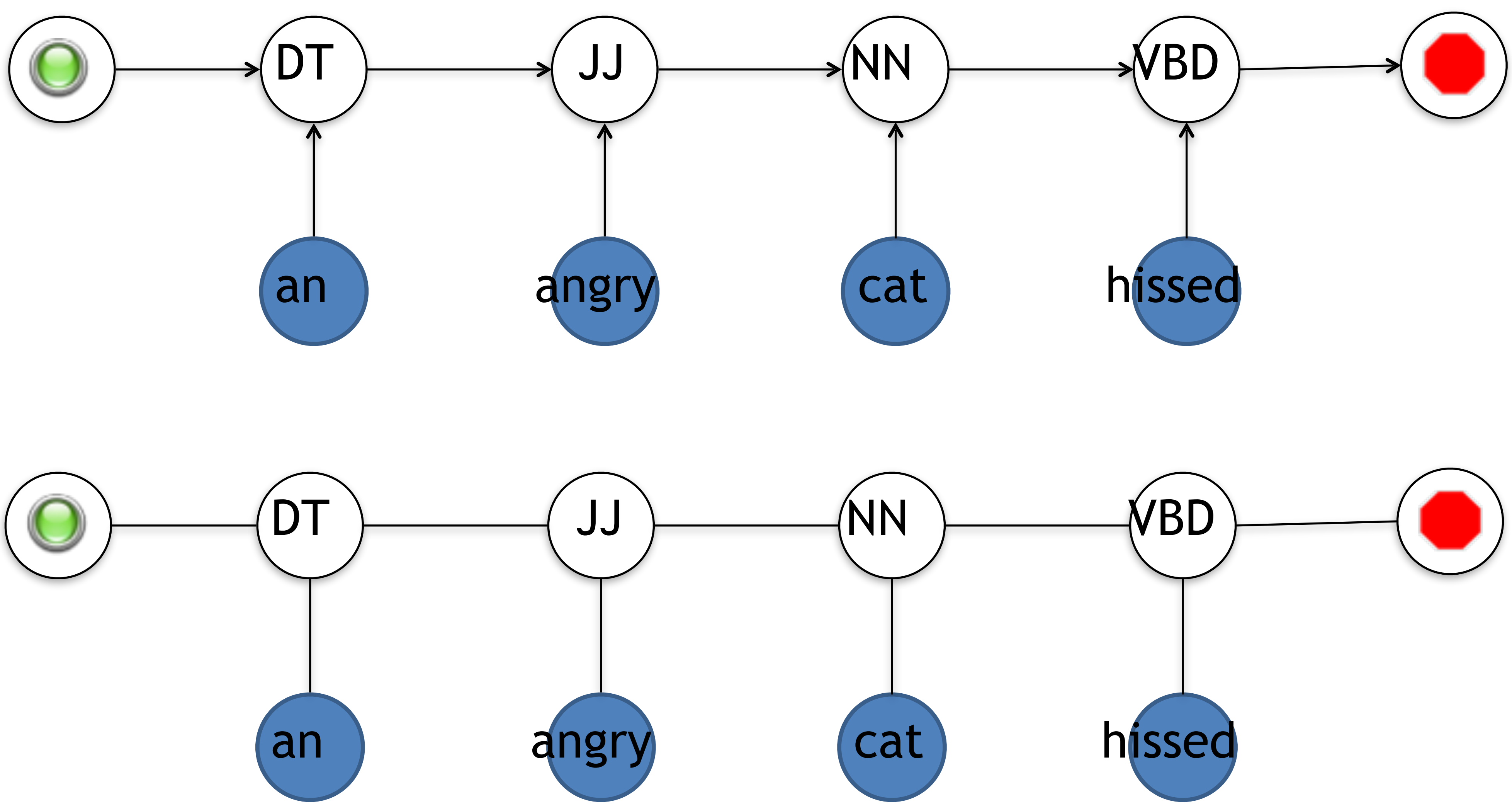
$$\frac{\partial \mathcal{L}}{\partial w_j} = \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in \mathcal{T}} \sum_{i=1}^{|\mathbf{x}|} \Big[f_j(y_i, \mathbf{x}, i, y_{i-1}) - \mathbb{E}_{p(y' | \mathbf{x}, i, y_{i-1}; \boldsymbol{w})} f_j(y', \mathbf{x}, i, y_{i-1}) \Big]$$

Conditional Random Fields

- Problems with MEMMs
  - What if we want to define a conditional distribution over trees? Or graphs? Or…?
  - Label bias
  - What if we want to define features like
    y_{-1} = DT & y_{+1} = VB

Solving Label Bias

- Intuitively, we would like each feature to contribute globally to the probability



Globally Normalized Models

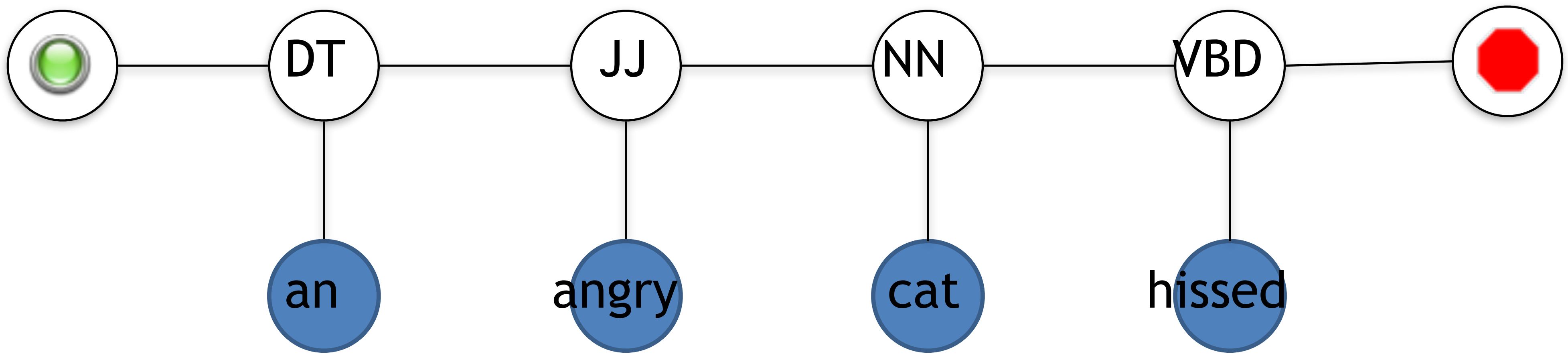
$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) = \frac{\exp \boldsymbol{w}^\top \boldsymbol{g}(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{y}' \in \mathcal{Y}_\mathbf{x}} \exp \boldsymbol{w}^\top \boldsymbol{g}(\mathbf{x}, \mathbf{y}')}$$

$$Z(\mathbf{x}; \boldsymbol{w}) = \sum_{\mathbf{y}' \in \mathcal{Y}_\mathbf{x}} \exp \boldsymbol{w}^\top \boldsymbol{g}(\mathbf{x}, \mathbf{y}')$$

Conditional Random Fields

- CRFs (Lafferty et al., 2001) are a special form of globally normalized models
  - They solve the label bias problem
  - They can be applied to arbitrary structures
  - They can use arbitrary features*
  - They generalize the notion of the logistic regression to cases where the output spaces has structure

CRFs for Sequence Labels



$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) = \frac{\exp \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{w}^\top \boldsymbol{f}(y_i, \mathbf{x}, i, y_{i-1})}{\sum_{\mathbf{y}' \in \Lambda^{|\mathbf{x}|}} \exp \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{w}^\top \boldsymbol{f}(y'_i, \mathbf{x}, i, y'_{i-1})}$$

Comparison to MEMMs

- CRF

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) = \frac{\exp \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{w}^\top \boldsymbol{f}(y_i, \mathbf{x}, i, y_{i-1})}{\sum_{\mathbf{y}' \in \Lambda^{|\mathbf{x}|}} \exp \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{w}^\top \boldsymbol{f}(y'_i, \mathbf{x}, i, y'_{i-1})}$$

- MEMM

$$p(y_i \mid \mathbf{x}, i, y_{i-1}; \boldsymbol{w}) = \frac{\exp \boldsymbol{w}^\top \boldsymbol{f}(y_i, \mathbf{x}, i, y_{i-1})}{\sum_{y' \in \Lambda} \exp \boldsymbol{w}^\top \boldsymbol{f}(y', \mathbf{x}, i, y_{i-1})}$$

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) = \prod_{i=1}^{|\mathbf{x}|} p(y_i \mid \mathbf{x}, i, y_{i-1}; \boldsymbol{w})$$

CRFs: Sum of their Parts

- A CRF is a globally normalized model in which **g** decomposes into local parts of the ***output*** structure

$$\Pi_i(\mathbf{x}, \mathbf{y}) = \langle y_i, \mathbf{x}, i, y_{i-1} \rangle$$

$$\boldsymbol{g}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{\#parts(\mathbf{x})} \boldsymbol{f}(\Pi_k(\mathbf{x}, \mathbf{y}))$$

Training CRFs

- Maximum likelihood estimation is straightforward, conceptually

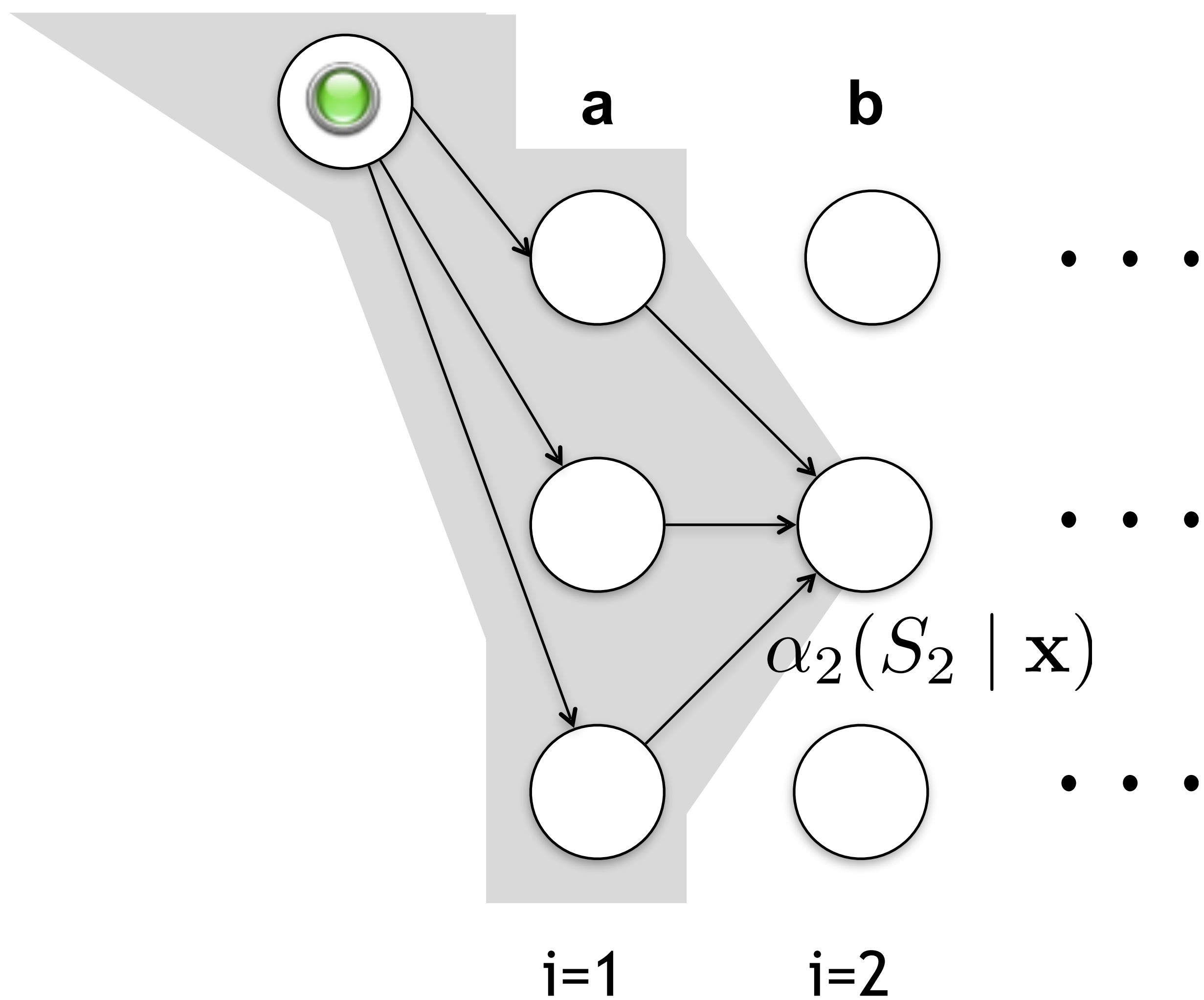
$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{w}) = \frac{\exp \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{w}^\top \boldsymbol{f}(y_i, \mathbf{x}, i, y_{i-1})}{\sum_{\mathbf{y}' \in \Lambda^{|\mathbf{x}|}} \exp \sum_{i=1}^{|\mathbf{x}|} \boldsymbol{w}^\top \boldsymbol{f}(y'_i, \mathbf{x}, i, y'_{i-1})}$$

$$\frac{\partial \mathcal{L}}{\partial w_j} = \sum_{i=1}^{\#parts(\mathbf{y})} \Big[\boldsymbol{f}(\Pi_i(\mathbf{x}, \mathbf{y})) - \mathbb{E}_{p(\mathbf{y}' \mid \mathbf{x}; \boldsymbol{w})} \boldsymbol{f}(\Pi_i(\mathbf{x}, \mathbf{y}')) \Big]$$

Efficient Inference

- If the parts factor into a sequence or a tree, then you can use polytime DP algorithms to
  - Solve for the MAP setting of Y
  - Compute the partition function
  - Compute posterior distributions over the settings of the variables in the parts

Forward Chart



$$\alpha_t(s \mid \mathbf{x}) = \sum_{r \to s} \alpha_{t-1}(r) \exp \boldsymbol{w}^\top \boldsymbol{f}(r, s, t, \mathbf{x})$$

A Word About Features

- Less "local" features require bigger part functions
  - This has a direct impact on the runtime of inference algorithms
  - But, in conditional models, you get to see the whole source "for free"
- Features are generally constructed by domain experts
  - They often have the form of templates %yi_suf(%xi)
- Feature learning or induction is becoming increasingly important
  - Conjunctions of basis features
  - Vector space ("distributed") representations