Conditional Random Fields

CSE 447 Lecture
Xiaochuang Han (Han)

Prerequisites

1. Sequence labeling task
   - Part-of-speech (POS) tagging

$$w_1, w_2, \ldots, w_n \longrightarrow y_1, y_2, \ldots, y_n$$

2. Logistic regression
   - Training and inference

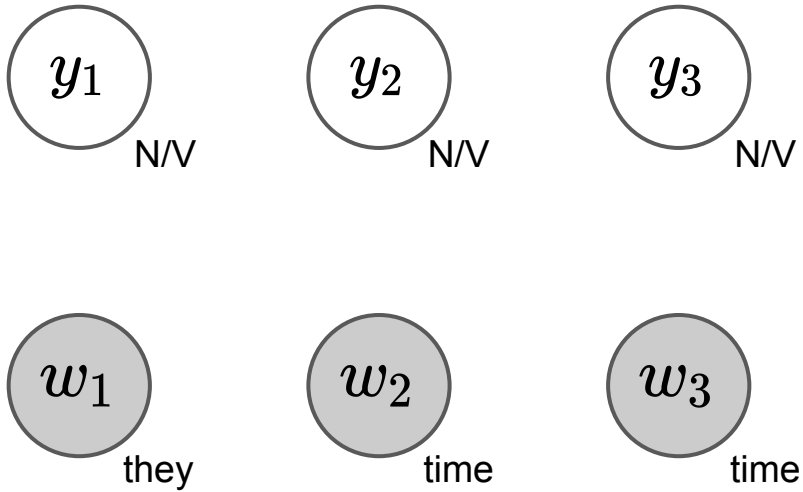
$$p_\theta(y \mid x)$$

3. Dynamic programming (optional)

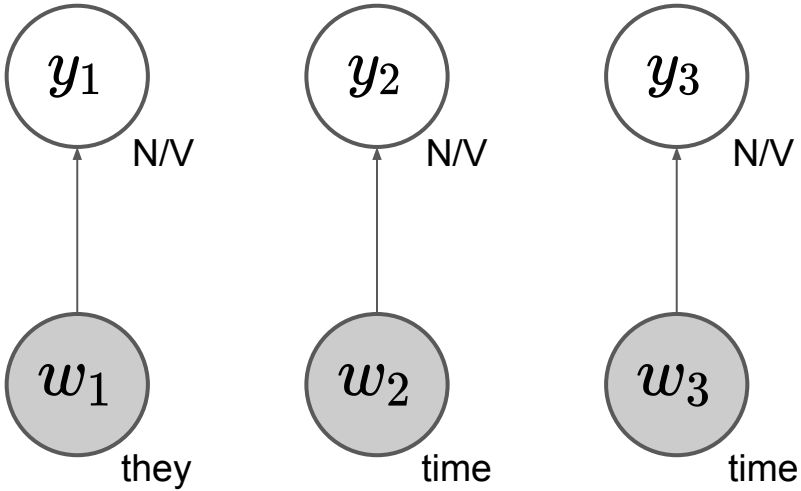
A POS-tagging example

POS?	*POS?*	*POS?*
they	time	time

A POS-tagging example

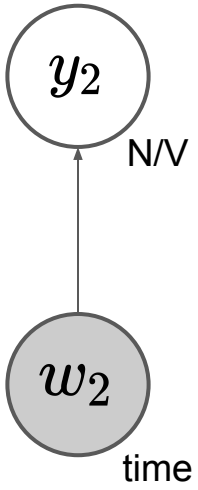


Logistic regression

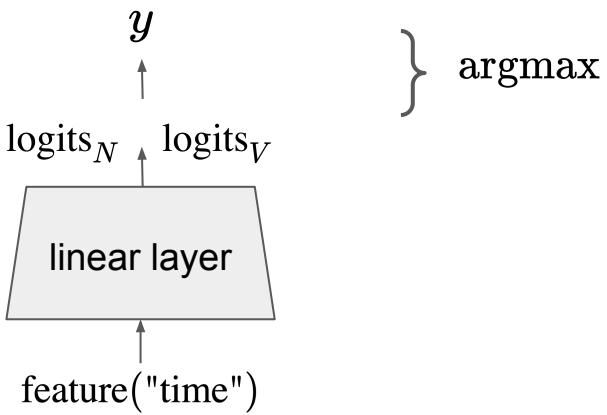


Predict each individual tag with logistic regression

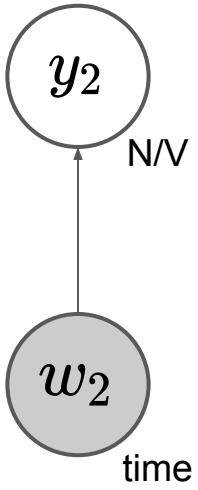
Logistic regression



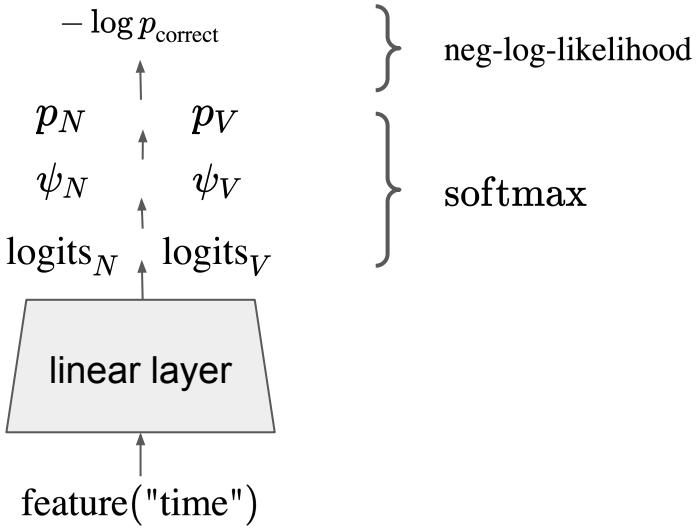
Inference



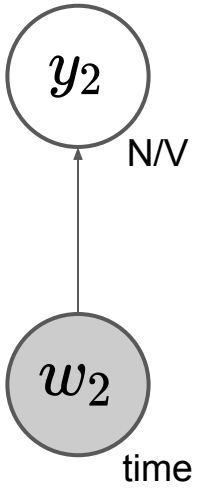
Logistic regression



Training



Logistic regression



Training

$-\log p_{\text{correct}}$ — neg-log-likelihood

p_N p_V

ψ_N ψ_V — softmax

logits_N logits_V

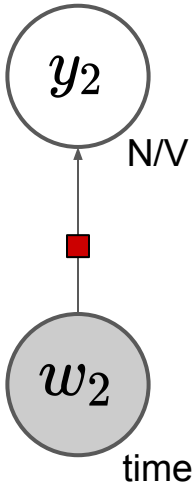
linear layer

feature("time")

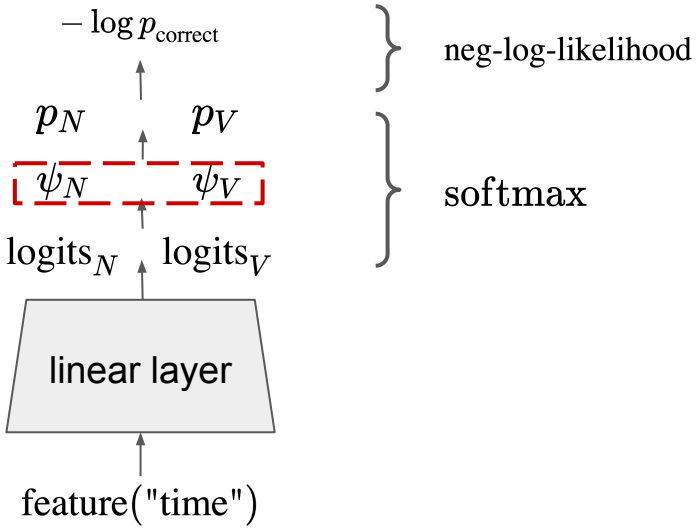
$\psi_N = \exp(\text{logits}_N)$

$p_N \propto \psi_N$ or more precisely, $p_N = \frac{\psi_N}{\psi_N + \psi_V}$ — softmax

Logistic regression



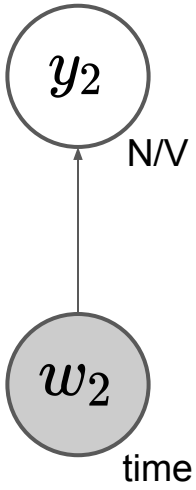
Training



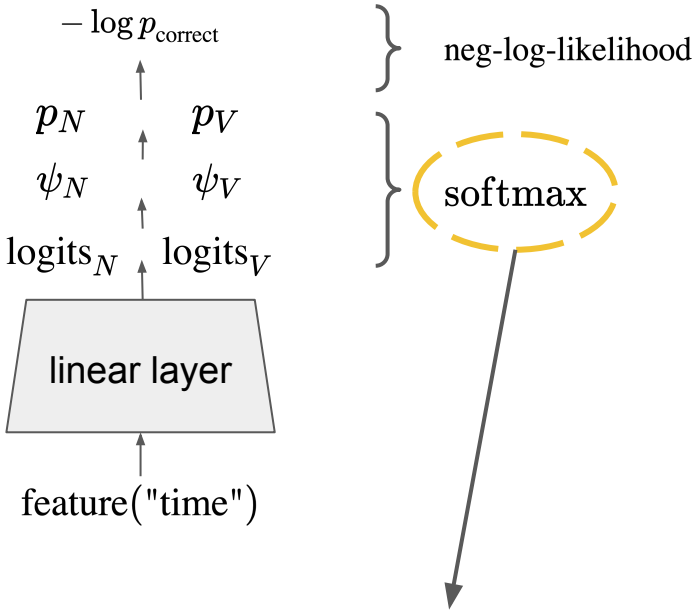
$$\psi_N = \exp(\text{logits}_N)$$

$p_N \propto \psi_N$ or more precisely, $p_N = \frac{\psi_N}{\psi_N + \psi_V}$

Logistic regression

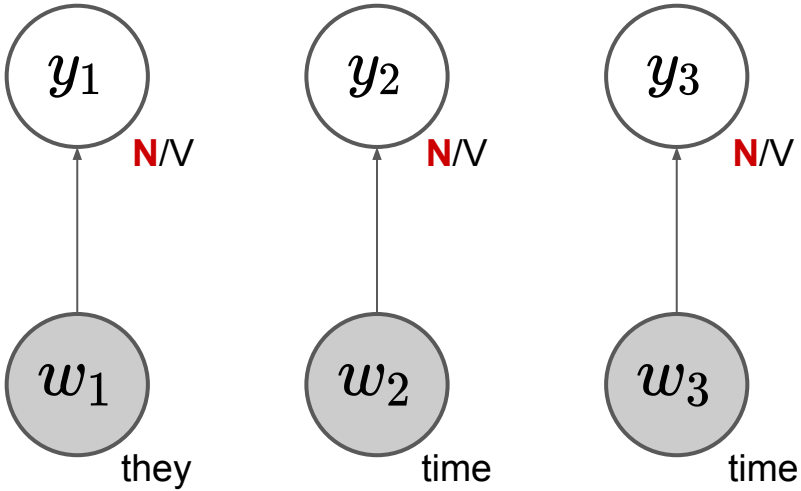


Training



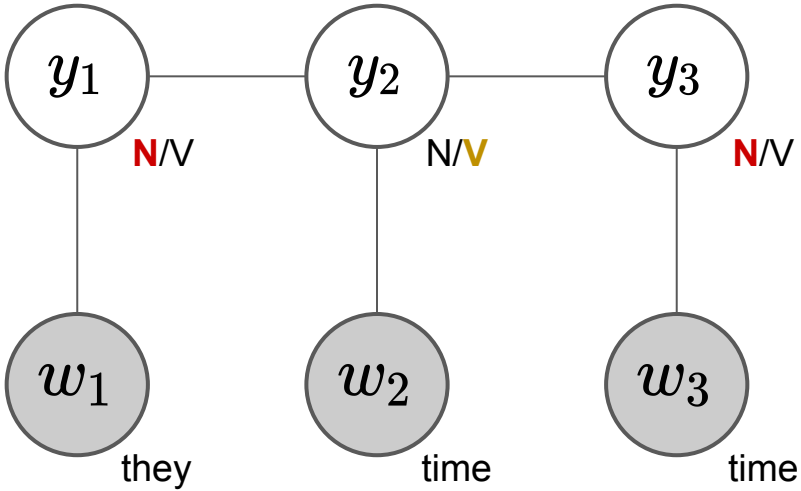
normalization is important but difficult in the sequence setup

Logistic regression



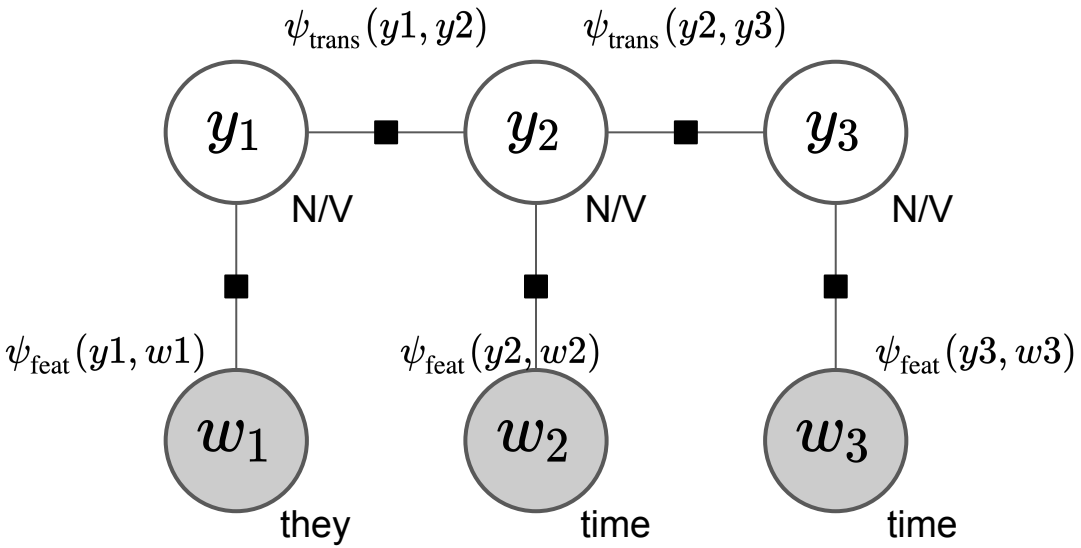
Predicting each individual tag with logistic regression is suboptimal

Conditional random fields



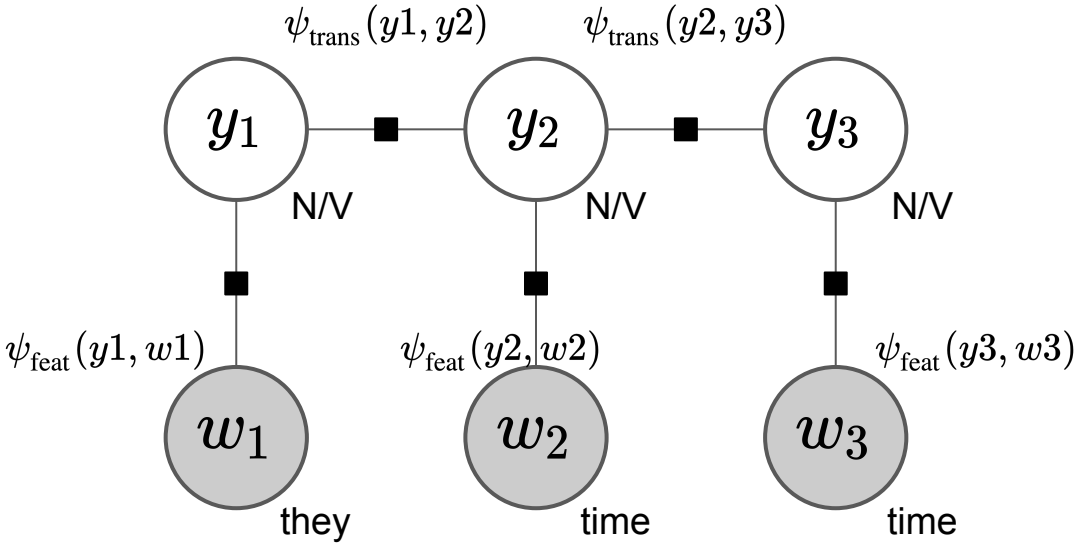
Incorporate structures between the labels

Conditional random fields



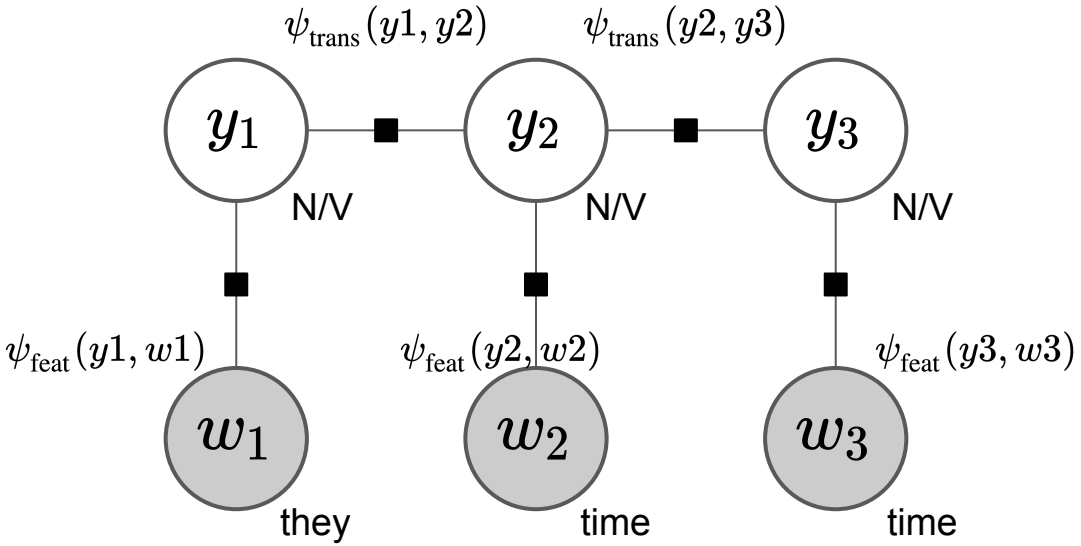
We define a series of scores ψ

Conditional random fields



These scores are similar to their counterparts in logistic regression: (0, +inf)

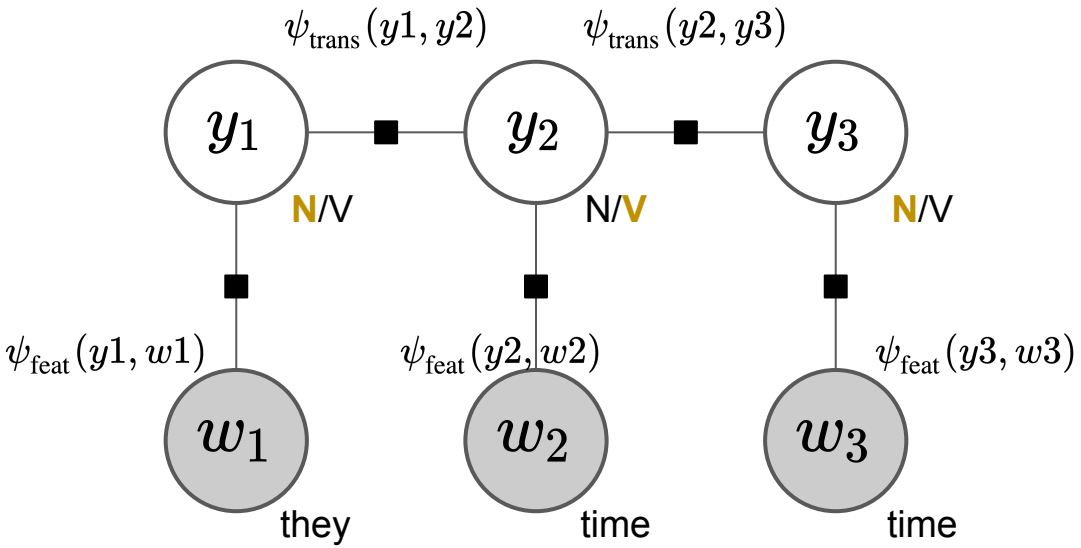
Conditional random fields



Again like in LR, these scores come from models with learnable parameters. In the homework:

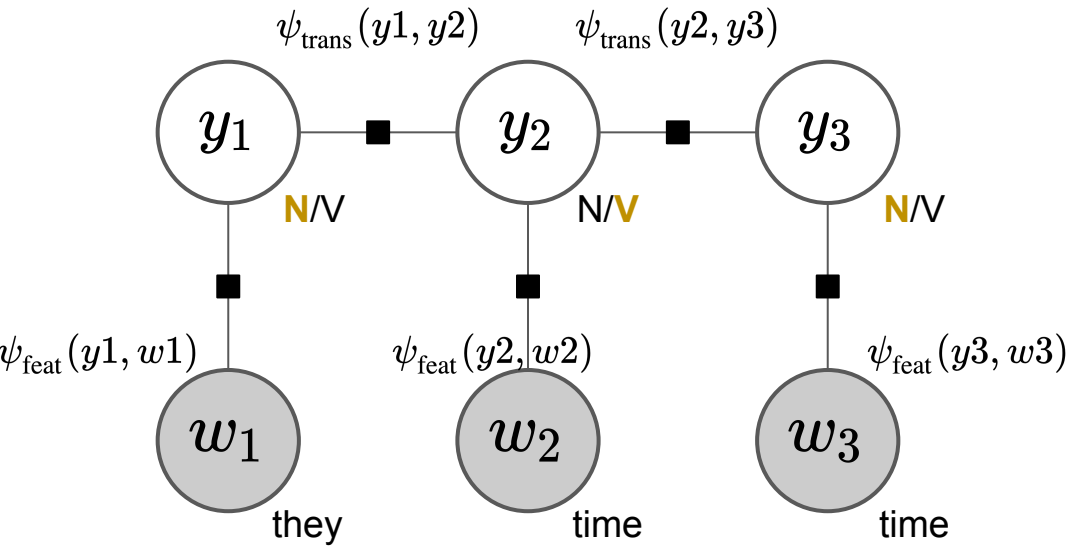
- ψ_{feat} is parameterized by a bidirectional LSTM
- ψ_{trans} is parameterized by a simple lookup table

Conditional random fields



The goal of training a CRF is to obtain the probability of the gold label sequence, and optimize the model parameters to maximize that probability.

Conditional random fields

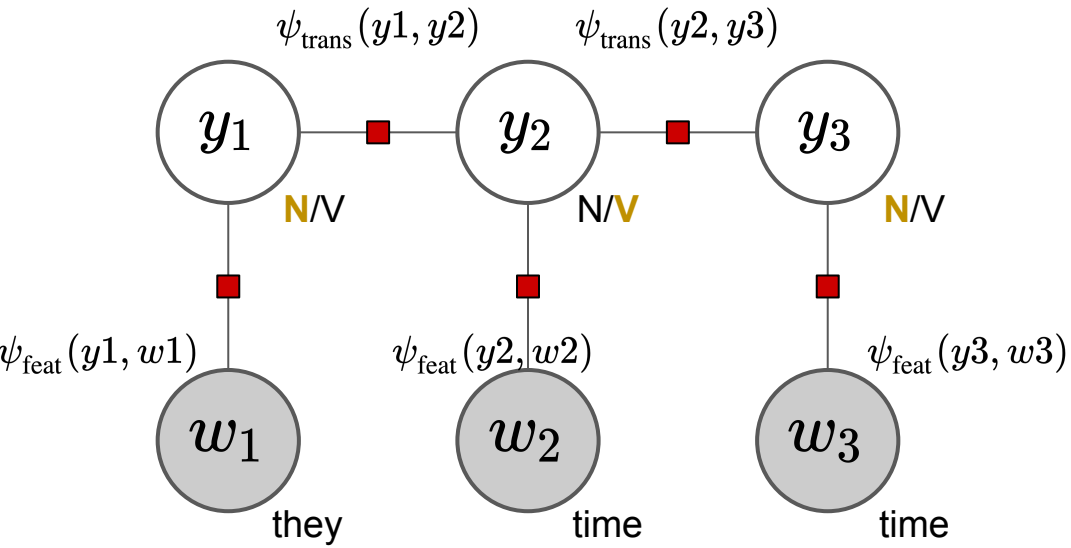


$$\psi(\cdot) \in (0, +\inf)$$

How to define the label sequence probability?

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

Conditional random fields



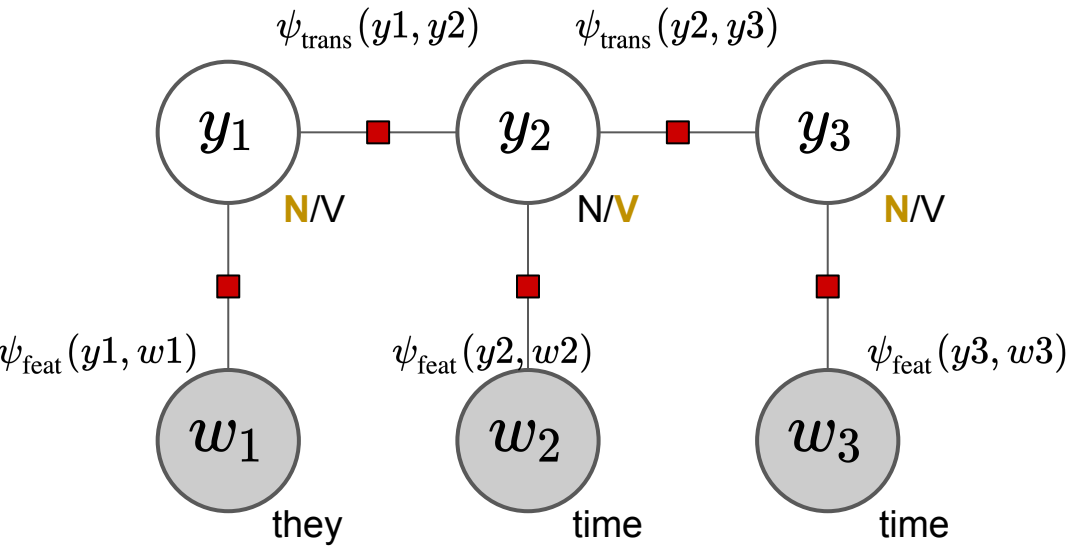
$$\psi(\cdot) \in (0, +\text{inf})$$

How to define the label sequence probability?

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

Conditional random fields



$$\psi(\cdot) \in (0, +\text{inf})$$

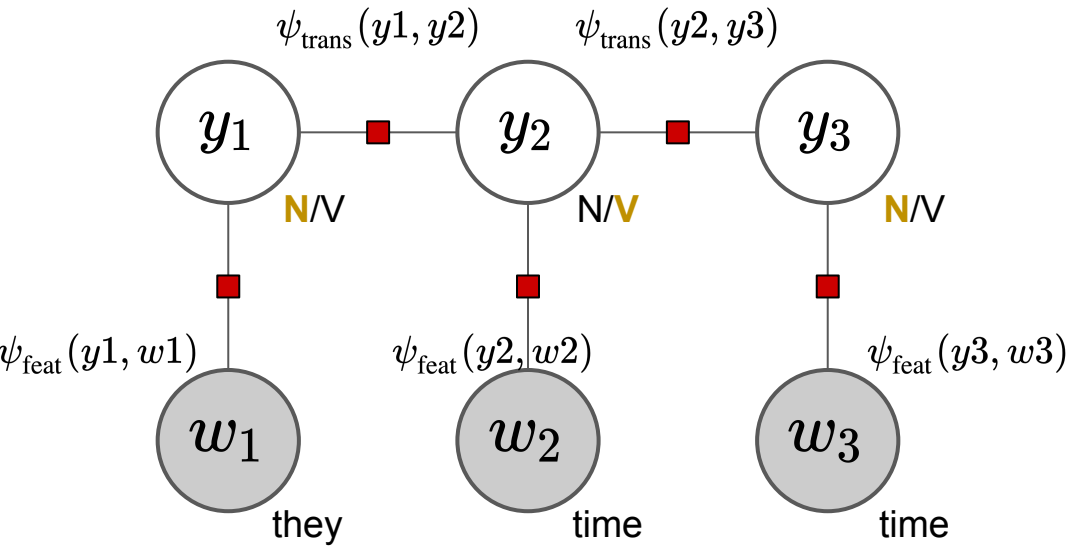
How to define the label sequence probability?

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

Conditional random fields



$$\psi(\cdot) \in (0, +\text{inf})$$

How to define the label sequence probability?

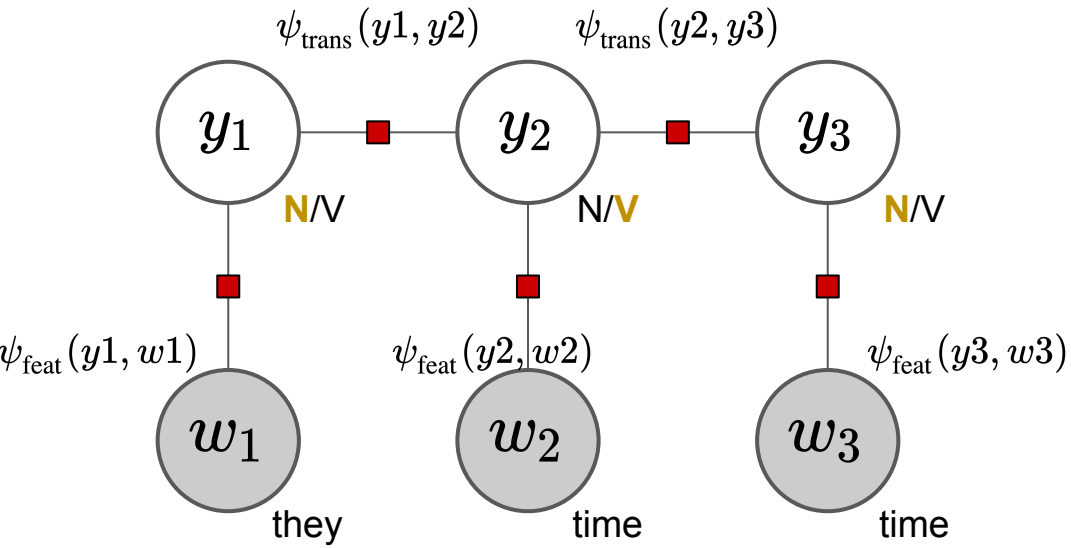
$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

We are essentially doing softmax here,
but the denominator is difficult to calculate!

Conditional random fields



$$\psi(\cdot) \in (0, +\text{inf})$$

How to define the label sequence probability?

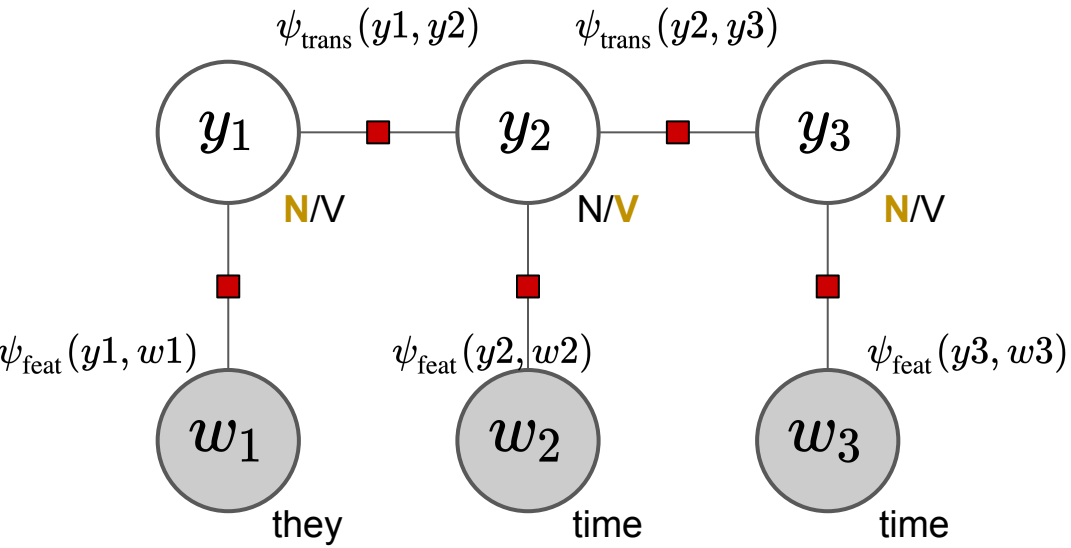
$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

We introduce the forward algorithm (a.k.a. sum-product algorithm) to obtain the denominator (a.k.a. partition function).

Conditional random fields



Alternative score definition here will result in intermediate exp or log transformations, but the algorithm stays the same.

$$\psi(\cdot) \in (0, +\inf)$$

How to define the label sequence probability?

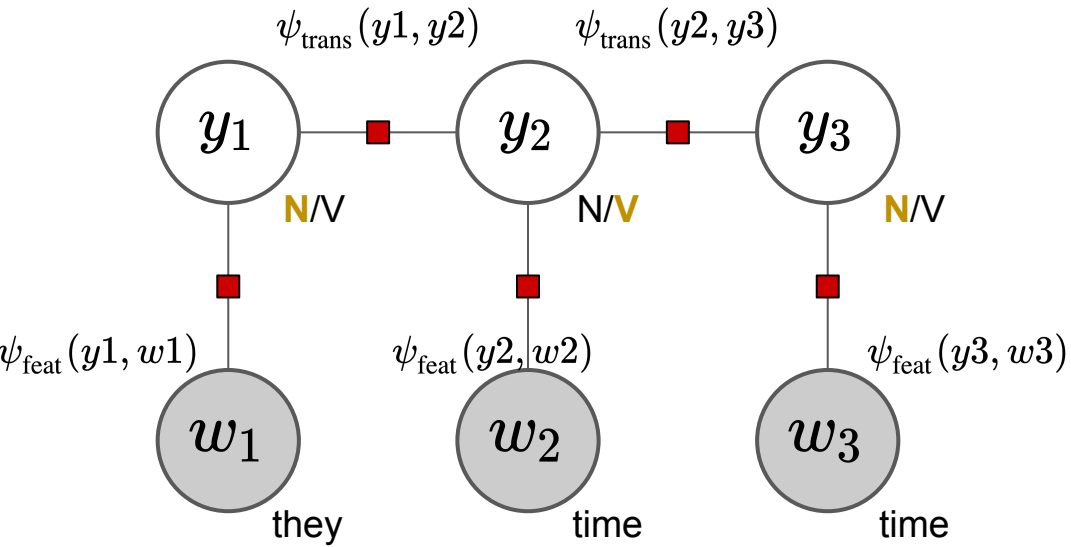
$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

We introduce the forward algorithm (a.k.a. sum-product algorithm) to obtain the denominator (a.k.a. partition function).

Conditional random fields



$$\psi(\cdot) \in (0, +\text{inf})$$

How to define the label sequence probability?

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

HW 6.2

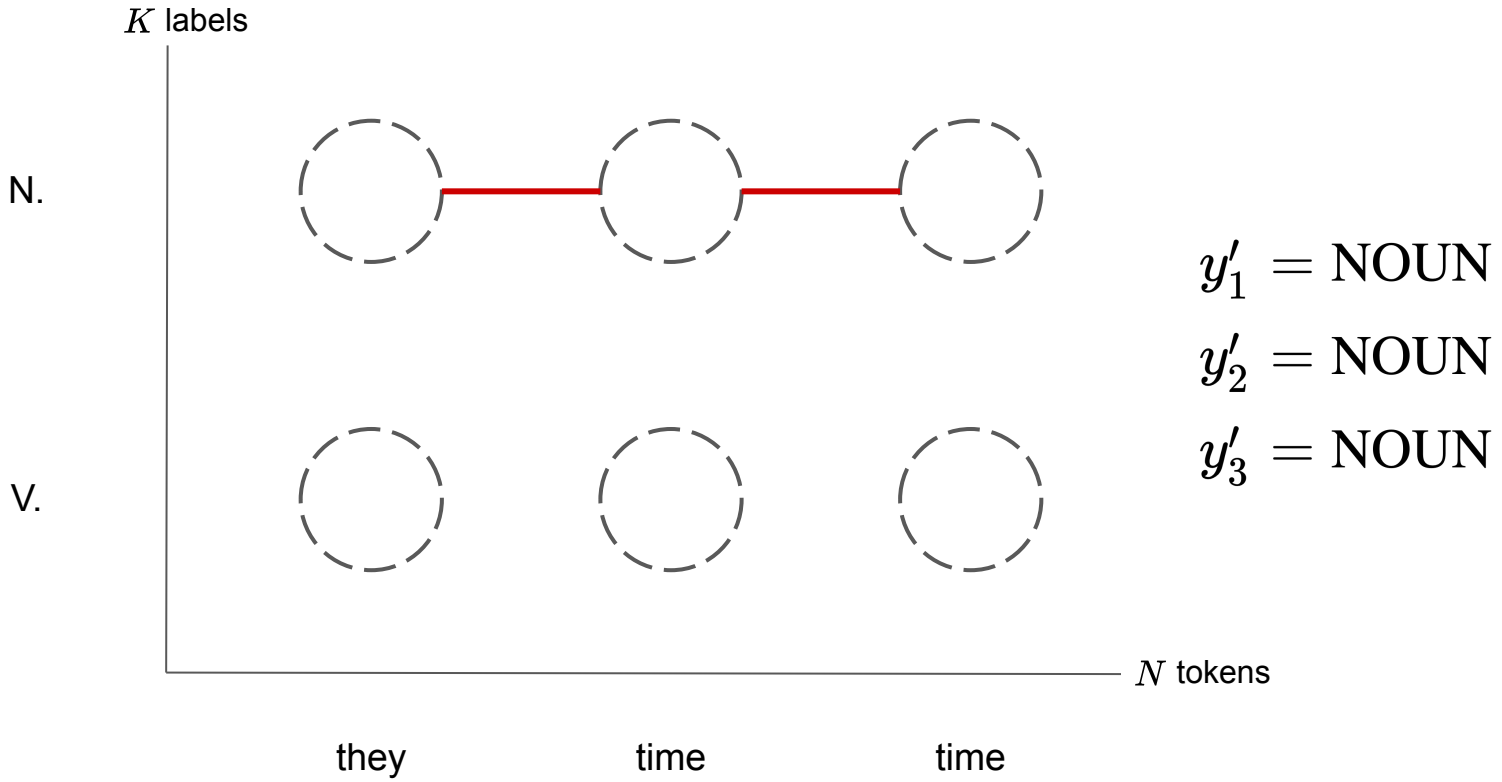
$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

HW 6.1

We introduce the forward algorithm (a.k.a. sum-product algorithm) to obtain the denominator (a.k.a. partition function).

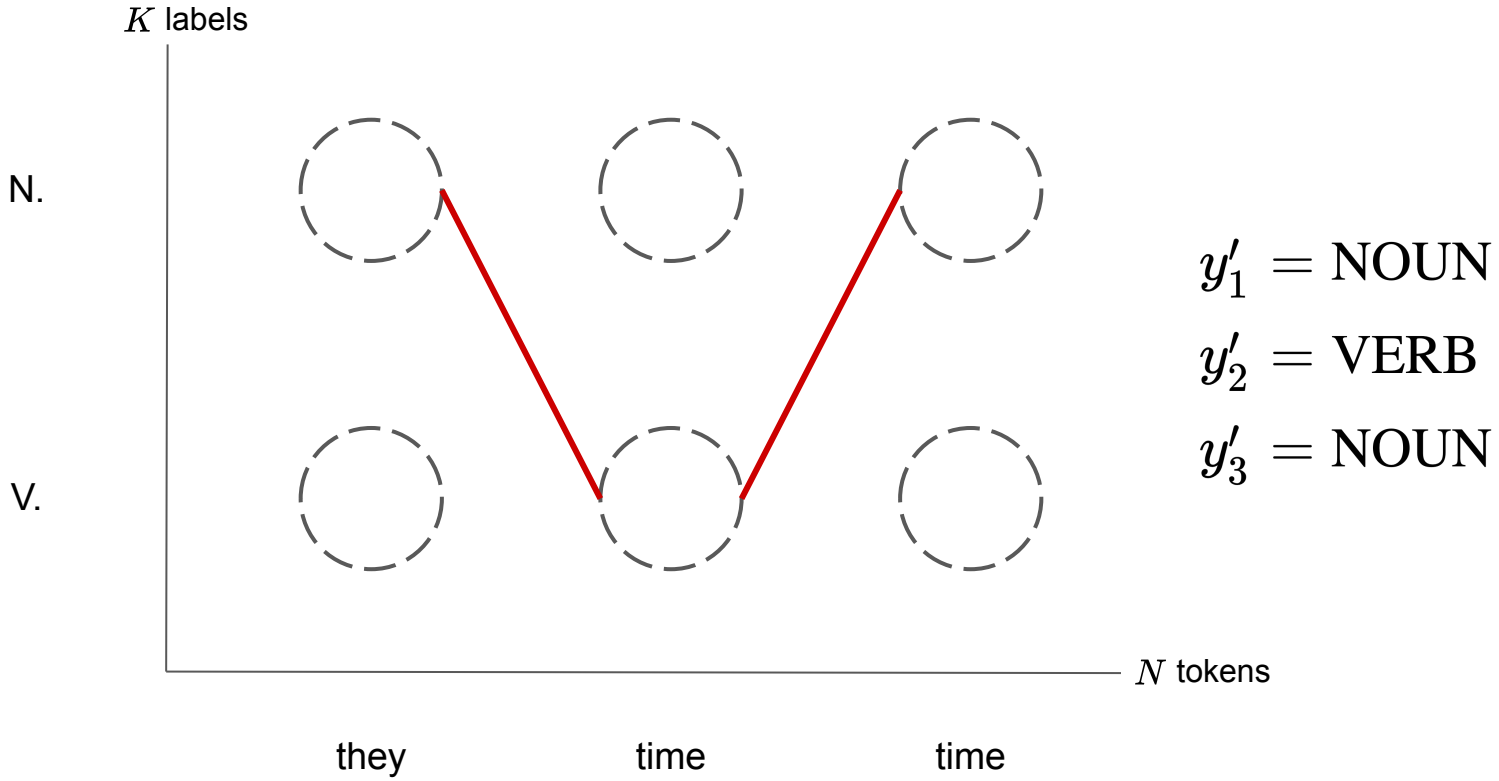
$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields



$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields

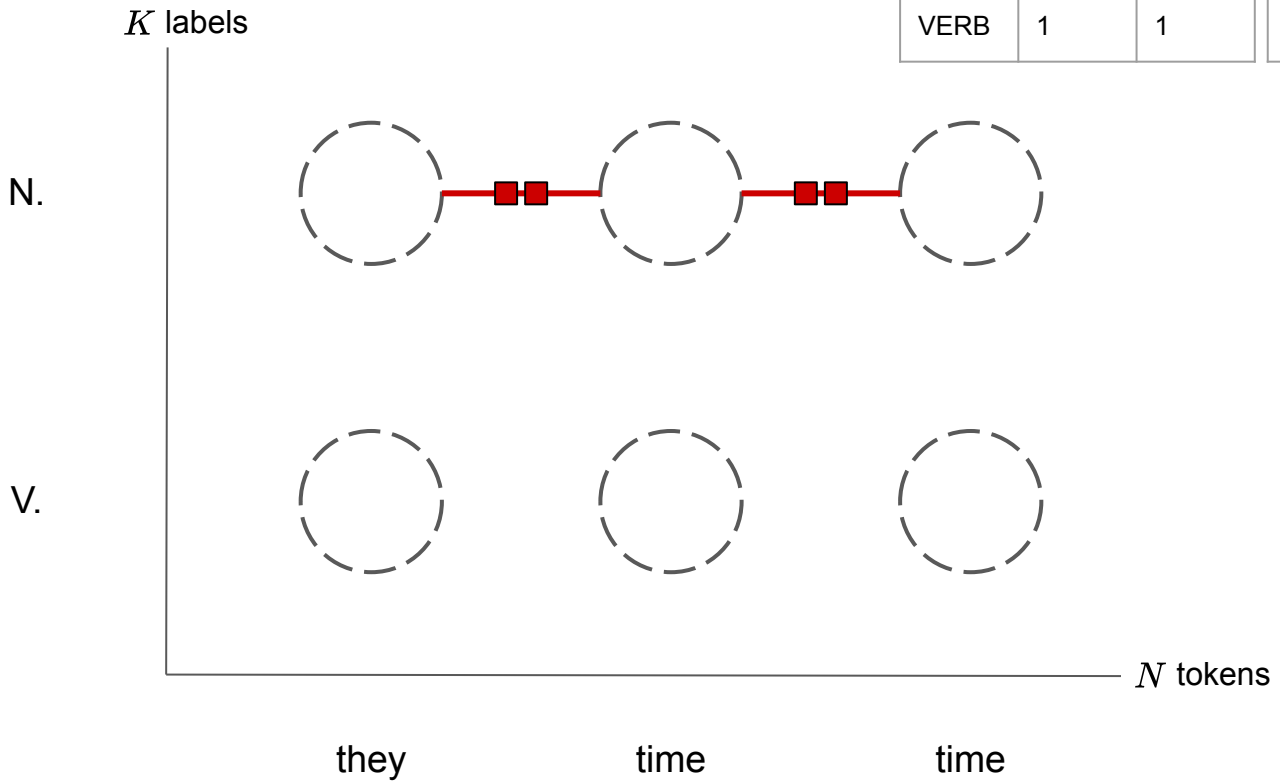


$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1

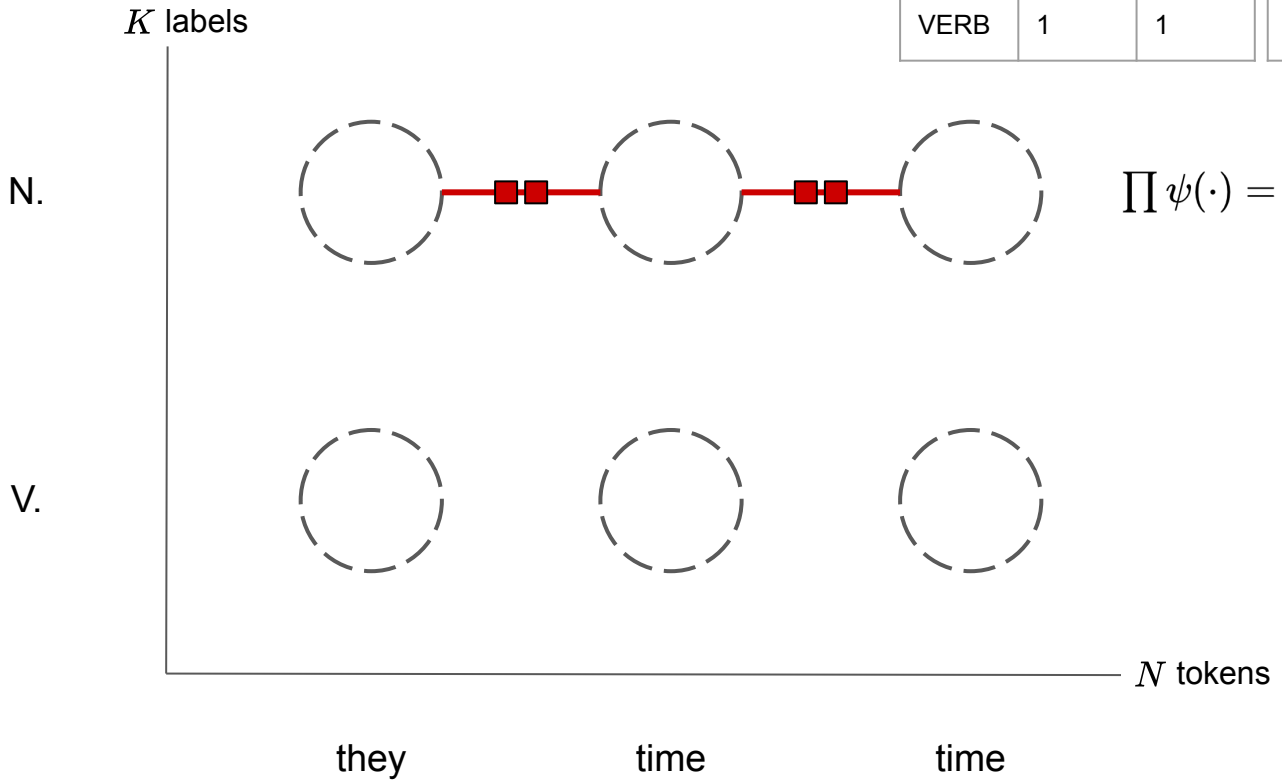


$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1



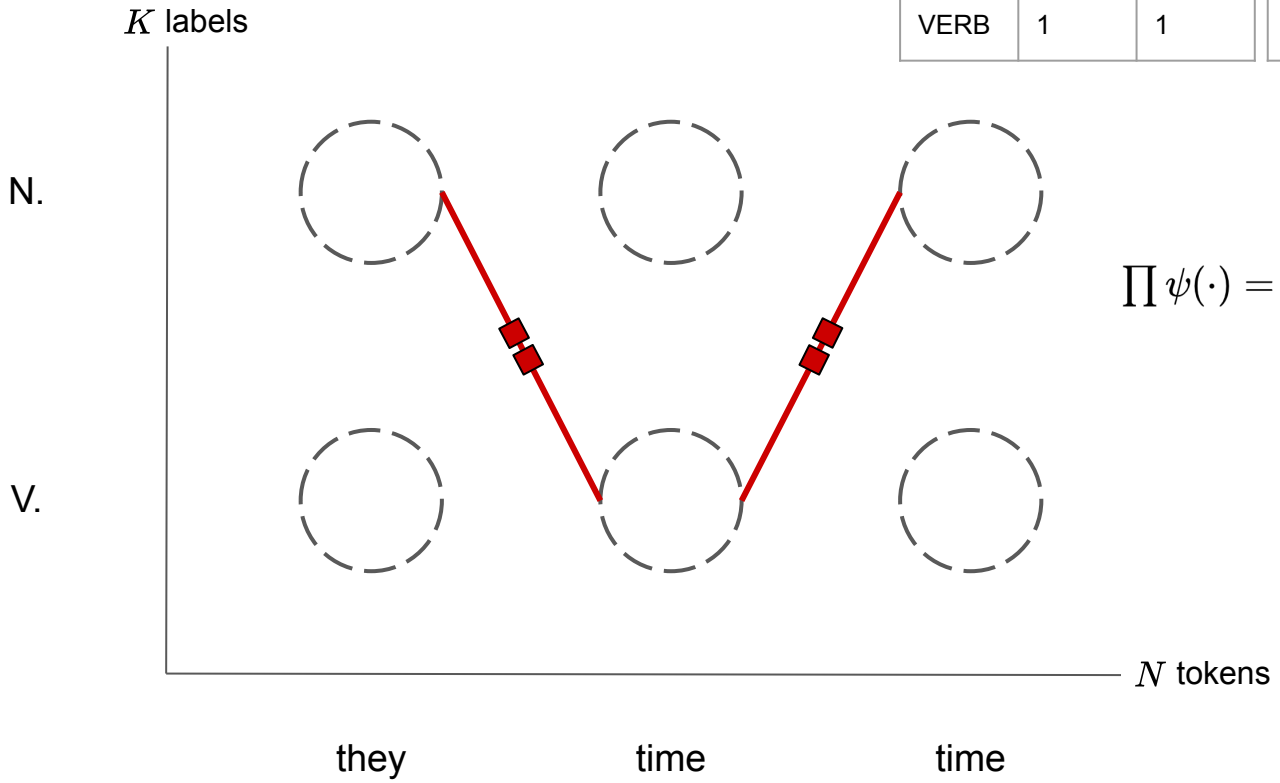
$$\prod \psi(\cdot) = 4 \times 2 \times 3 \times 2 \times 3$$

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1



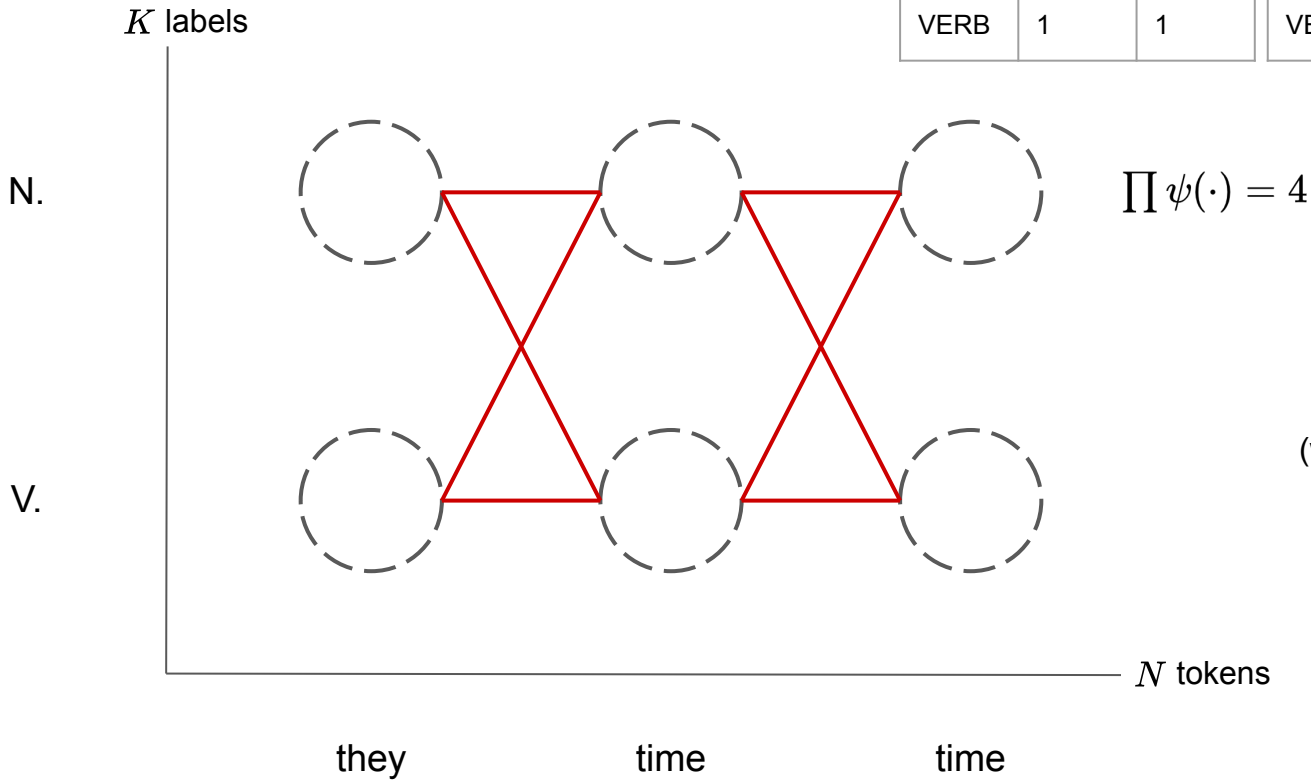
$$\prod \psi(\cdot) = 4 \times 3 \times 1 \times 3 \times 3$$

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1



$$\prod \psi(\cdot) = 4 \times 2 \times 3 \times 2 \times 3$$

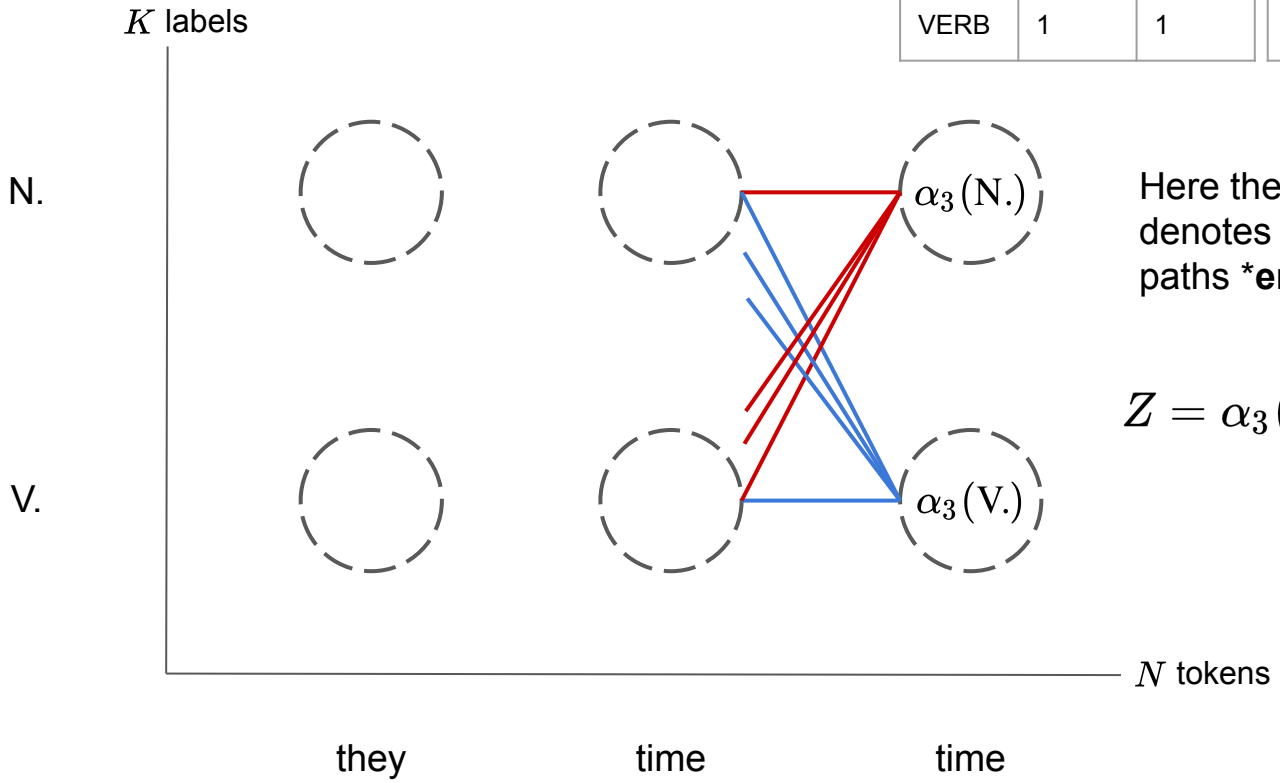
(we want their sum)

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1



Here the alpha variable denotes the sum of all paths *__ending__* at the cell.

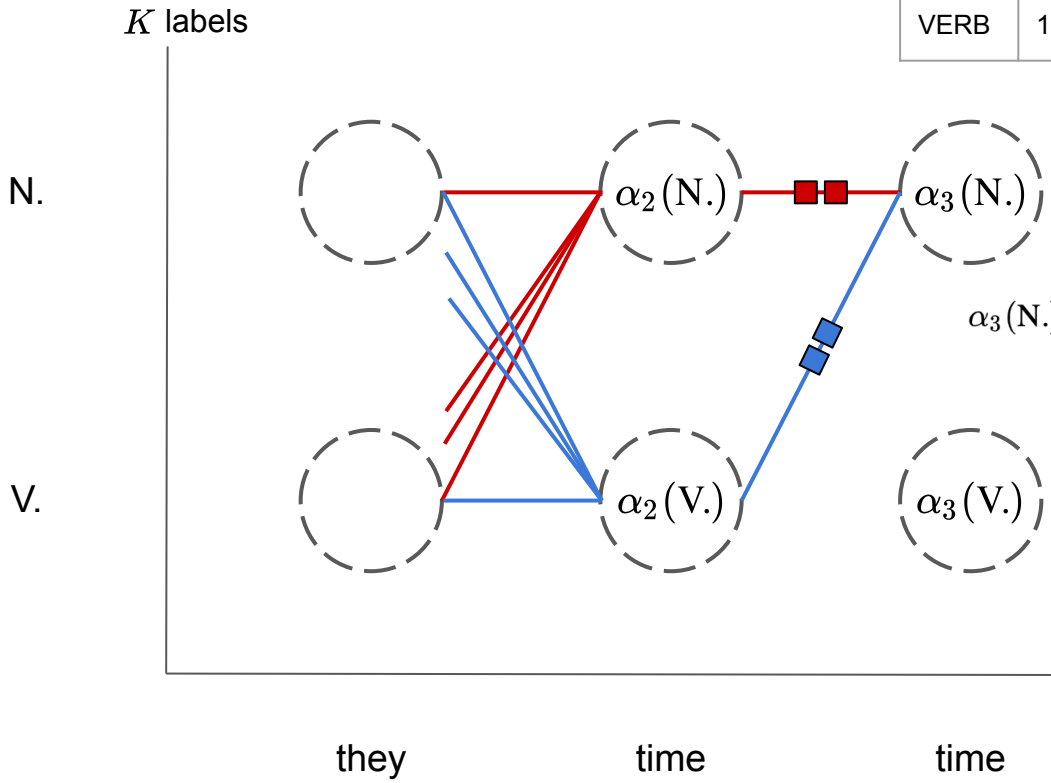
$$Z = \alpha_3(\text{N.}) + \alpha_3(\text{N.})$$

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

Conditional random fields

ψ_{feat}	"they"	"time"
NOUN	4	3
VERB	1	1

ψ_{trans}	NOUN	VERB
NOUN	2	3
VERB	3	1



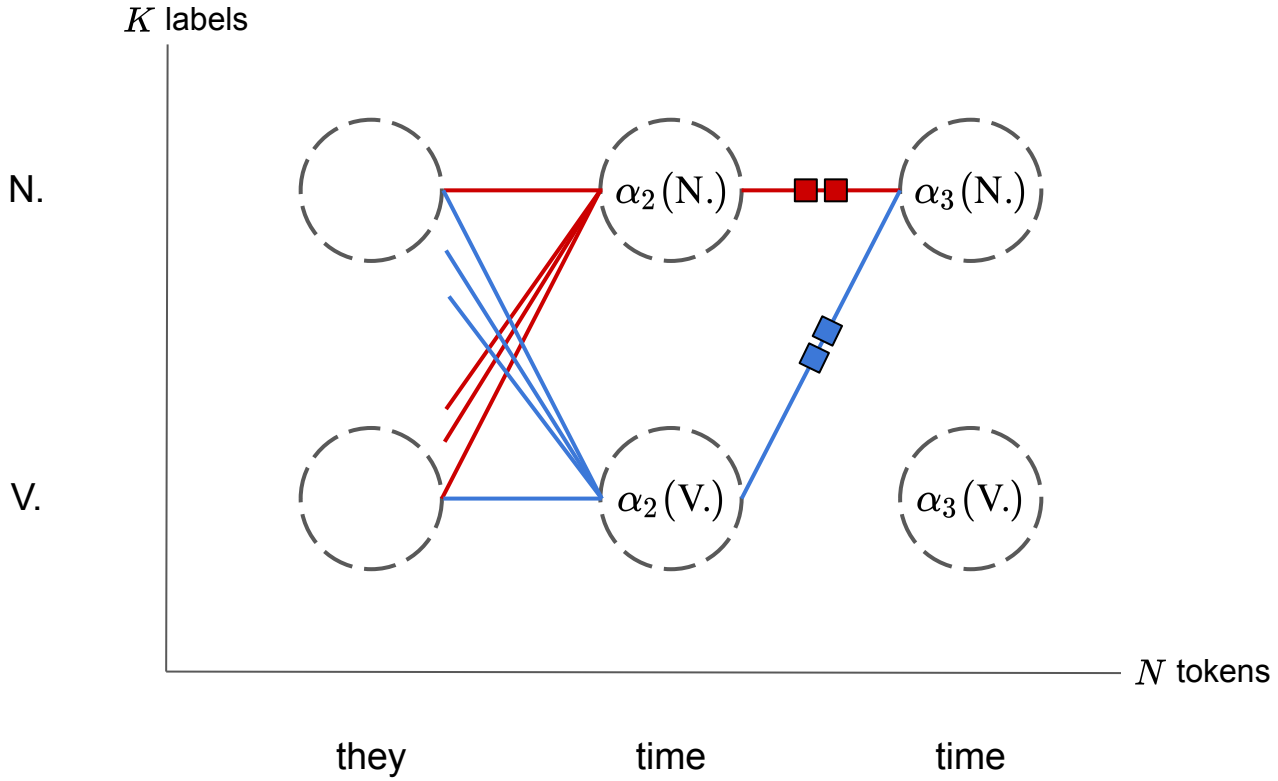
Stepping backward, we can compute each of the alpha variable recursively.

$$\alpha_3(\text{N.}) = \alpha_2(\text{N.})\psi_{\text{trans}}(\text{N.}, \text{N.})\psi_{\text{feat}}(\text{N.}, \text{"time"}) + \alpha_2(\text{V.})\psi_{\text{trans}}(\text{V.}, \text{N.})\psi_{\text{feat}}(\text{N.}, \text{"time"})$$

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

$$Z = \sum_{y_N} \alpha_N(y_N)$$

Conditional random fields



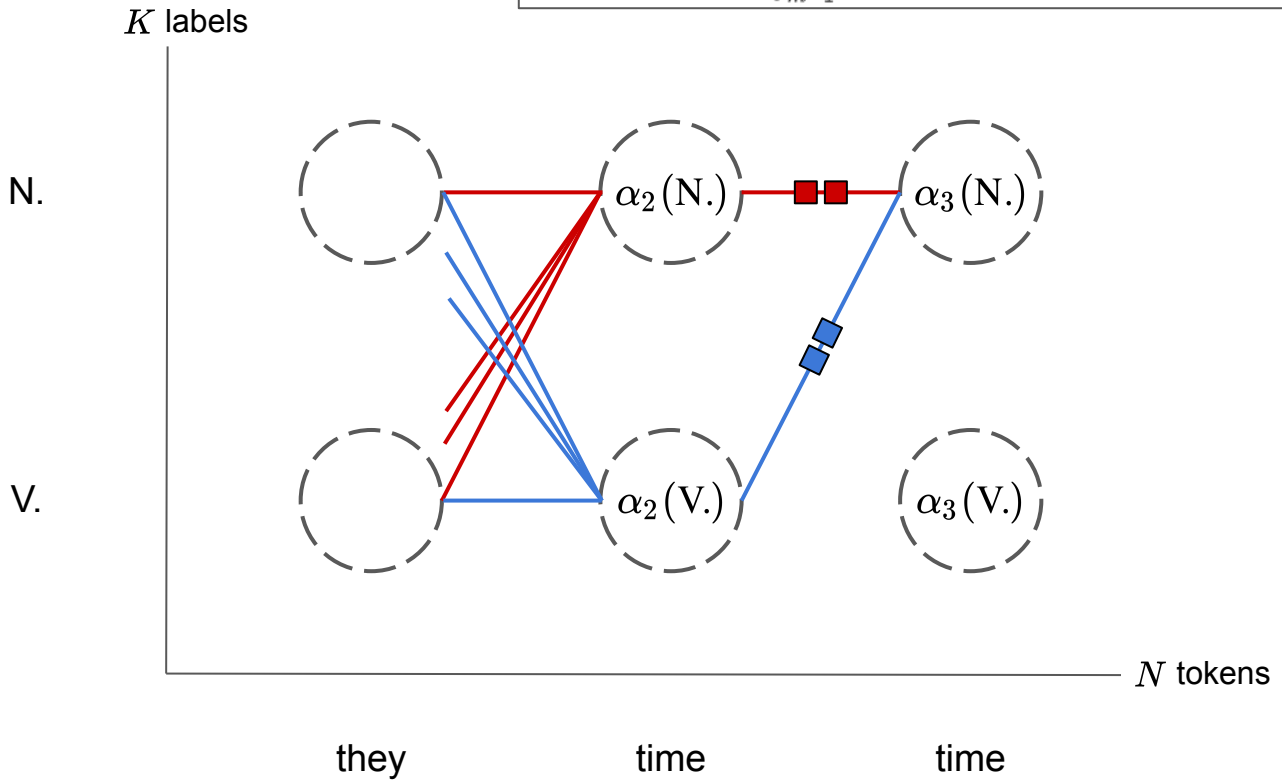
Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

$$Z = \sum_{y_N} \alpha_N(y_N)$$

$$\alpha_m(y_m) = \sum_{y_{m-1}} \alpha_{m-1}(y_{m-1})\psi_{\text{trans}}(y_{m-1}, y_m)\psi_{\text{feat}}(y_m, w_m)$$

for $m = 1, 2, \ldots, N$



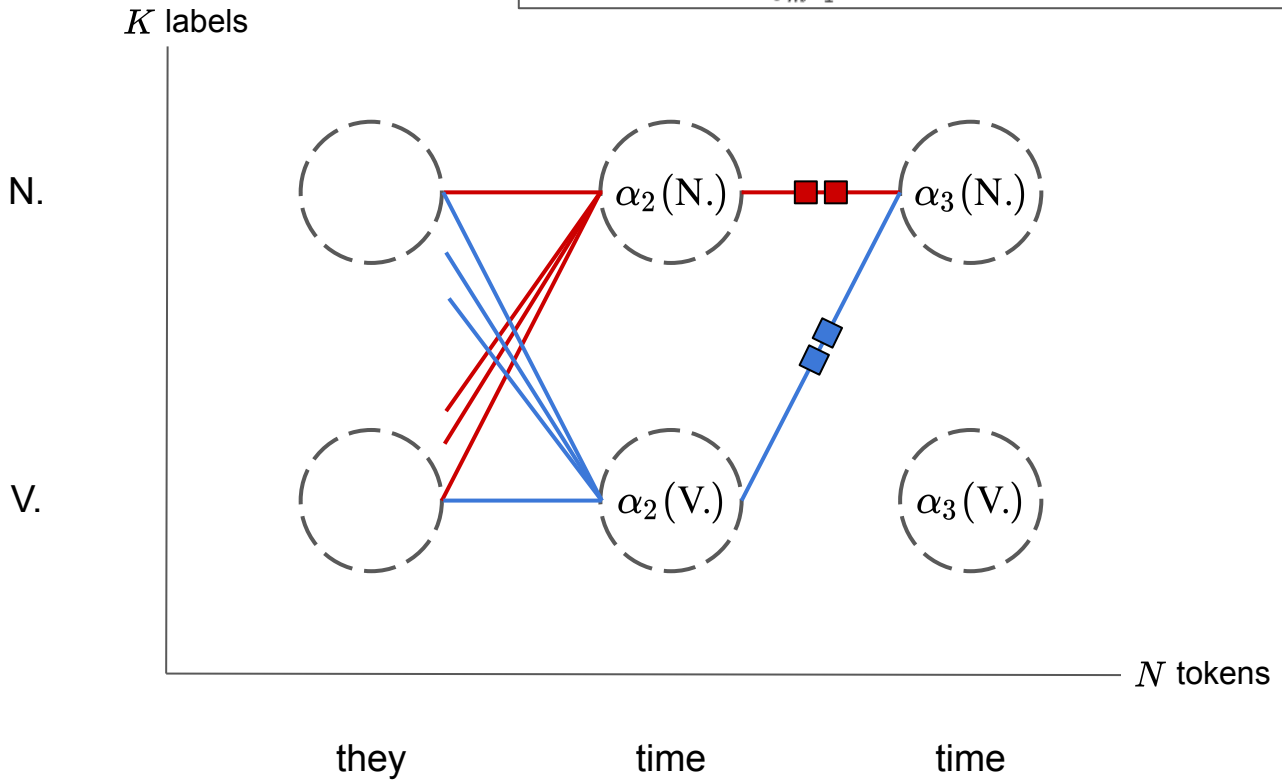
$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n) \psi_{\text{trans}}(y'_n, y'_{n+1})$$

$$Z = \sum_{y_N} \alpha_N(y_N)$$

[TODO] manual derivation?

$$\alpha_m(y_m) = \sum_{y_{m-1}} \alpha_{m-1}(y_{m-1}) \psi_{\text{trans}}(y_{m-1}, y_m) \psi_{\text{feat}}(y_m, w_m)$$

for $m = 1, 2, \ldots, N$



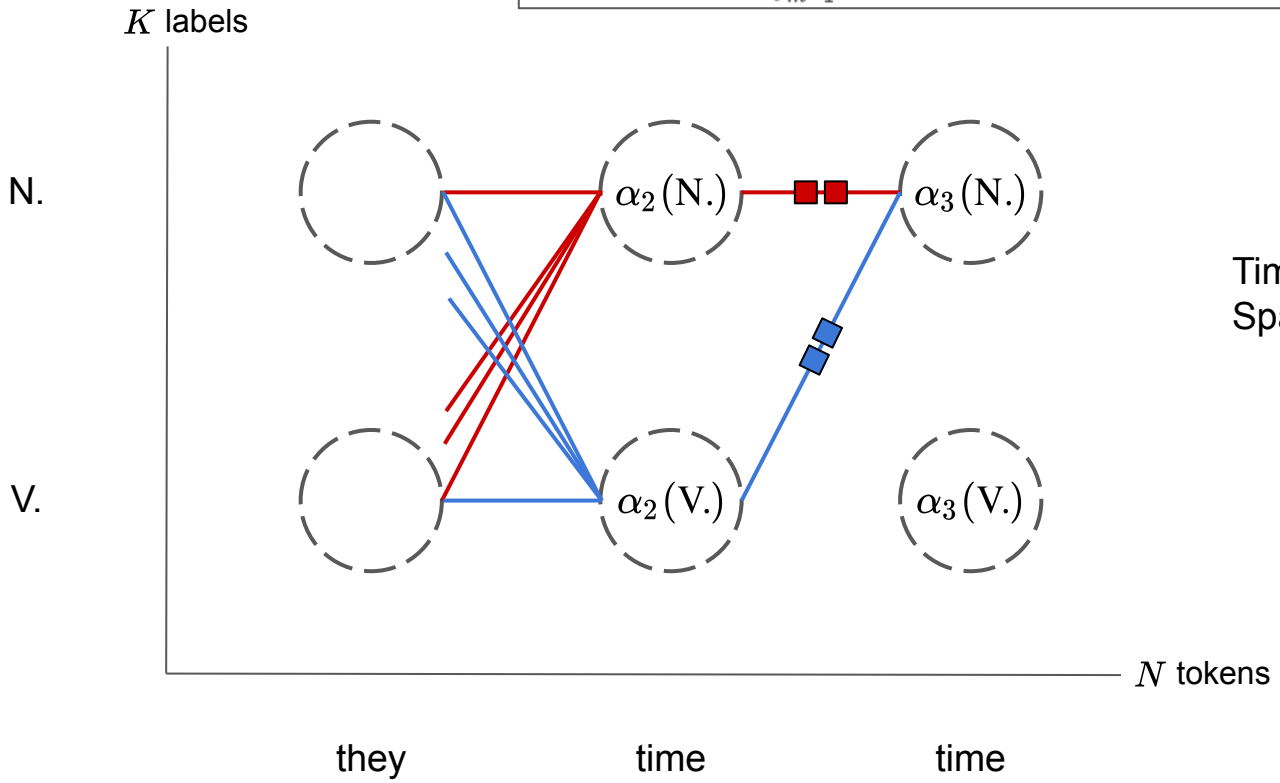
Conditional random fields

$$Z = \sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})$$

$$Z = \sum_{y_N} \alpha_N(y_N)$$

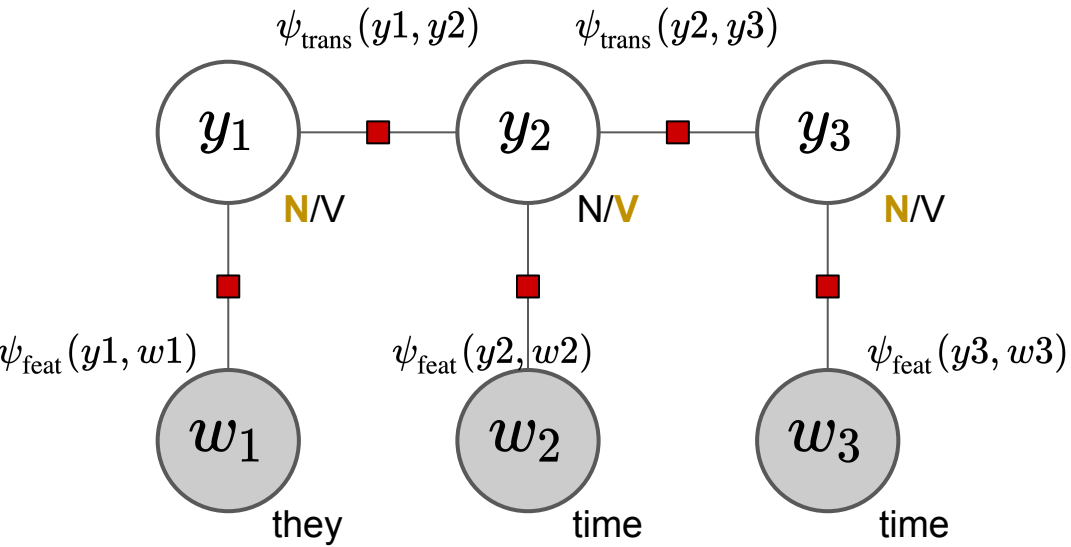
$$\alpha_m(y_m) = \sum_{y_{m-1}} \alpha_{m-1}(y_{m-1})\psi_{\text{trans}}(y_{m-1}, y_m)\psi_{\text{feat}}(y_m, w_m)$$

for $m = 1, 2, \ldots, N$



Time complexity?
Space complexity?

Conditional random fields



$$\psi(\cdot) \in (0, +\text{inf})$$

How to define the label sequence probability?

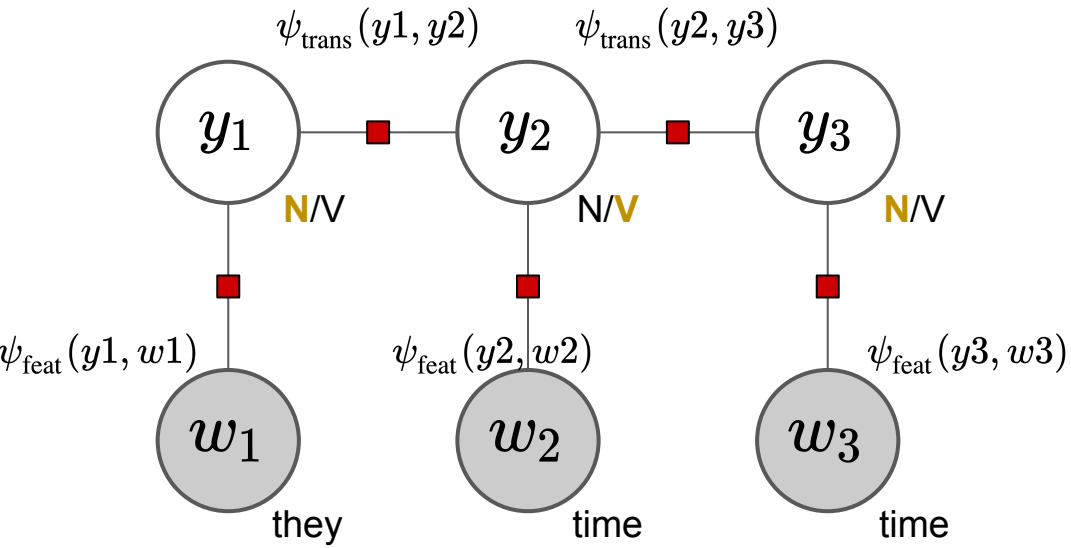
$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w}) \propto \prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})$$

$$= \frac{\prod_{n=1}^{N} \psi_{\text{feat}}(y_n, w_n)\psi_{\text{trans}}(y_n, y_{n+1})}{\sum_{\boldsymbol{y}' \in \mathcal{Y}(\boldsymbol{w})} \prod_{n=1}^{N} \psi_{\text{feat}}(y'_n, w_n)\psi_{\text{trans}}(y'_n, y'_{n+1})}$$

$$= \text{something computable!!!}$$

Conditional random fields



$$\psi(\cdot) \in (0, +\text{inf})$$

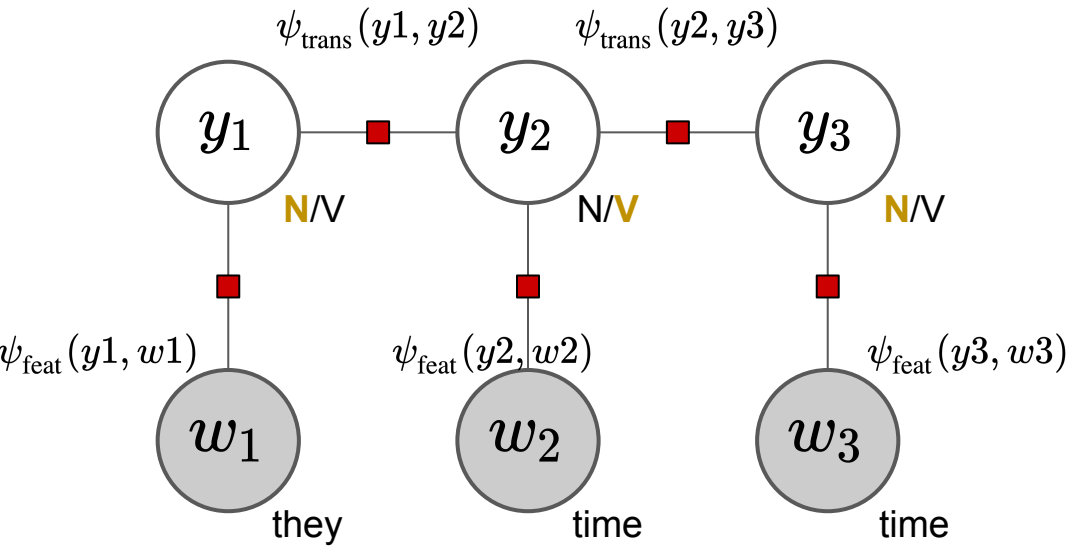
How to maximize the gold sequence probability?

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{w})$$

$$\theta \leftarrow \theta - \eta \nabla_\theta (-\log p_\theta(\boldsymbol{y} \mid \boldsymbol{w}))$$

Gradient descent, or any of your favorite optimizers :)

Conditional random fields



$$\psi(\cdot) \in (0, +\text{inf})$$

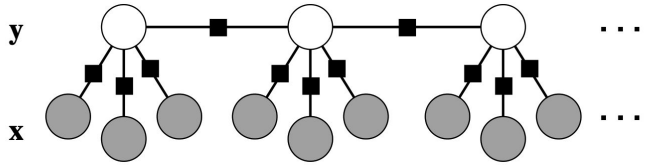
How to do inference on test-time inputs with the learned model?

The Viterbi algorithm, a.k.a. max-product algorithm
(This part is the same as HMMs)

Further details can be found in Chapter 7 of the Eisenstein textbook here.

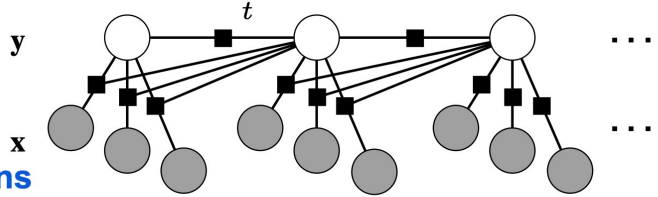
Other types of CRF

Direct Extension of HMM



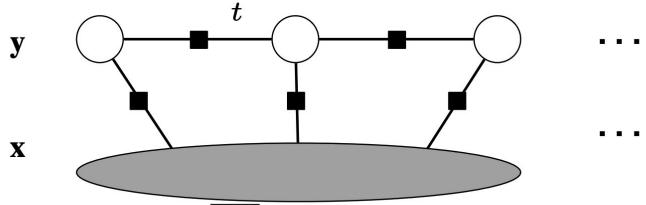
$$p(y \mid x) \propto \prod_t \psi_t(y_t, x_t)\psi_{t,t+1}(y_t, y_{t+1})$$

State Transitions depend on Observations



$$p(y \mid x) \propto \prod_t \psi_t(y_t, x_t)\psi_{t,t+1}(y_t, y_{t+1}, x_t)$$

Arbitrary Non-Local Features



$$p(y \mid x) \propto \prod_t \psi_t(y_t, x)\psi_{t,t+1}(y_t, y_{t+1}, x)$$

source