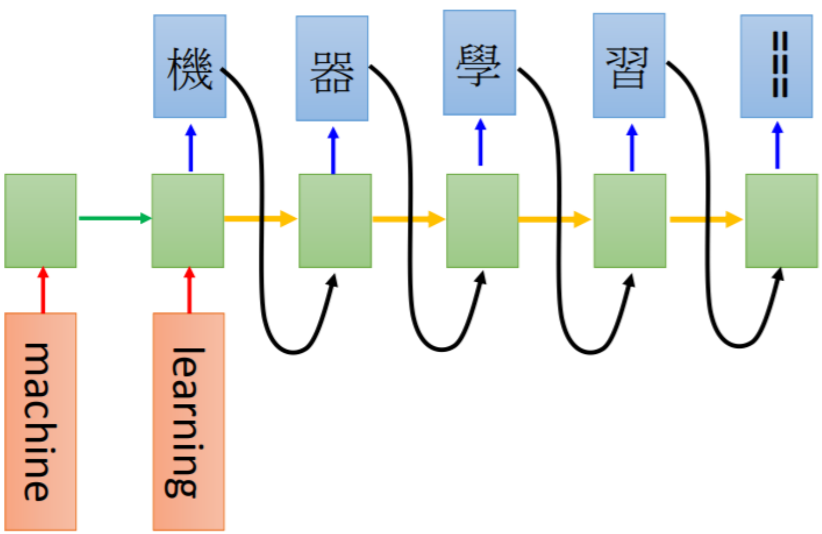
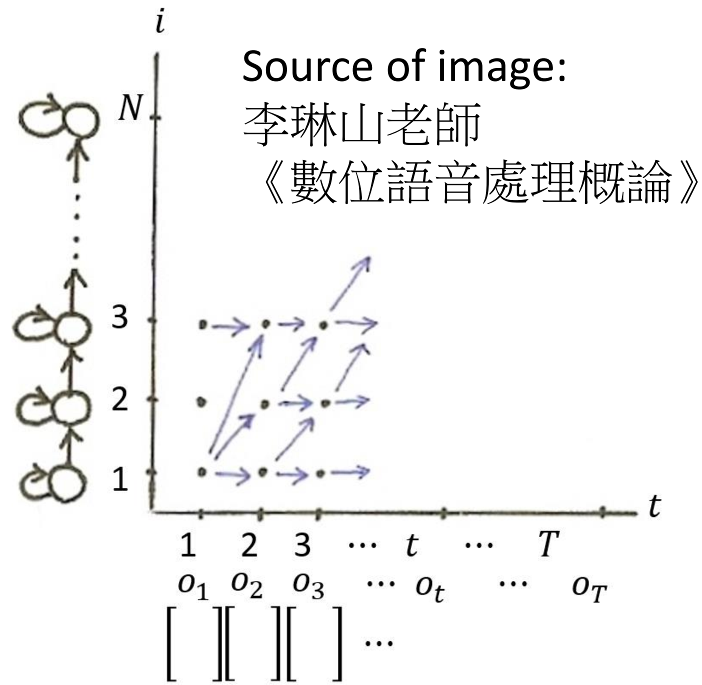
Speech Recognition

HUNG-YI LEE 李宏毅

Two Points of Views

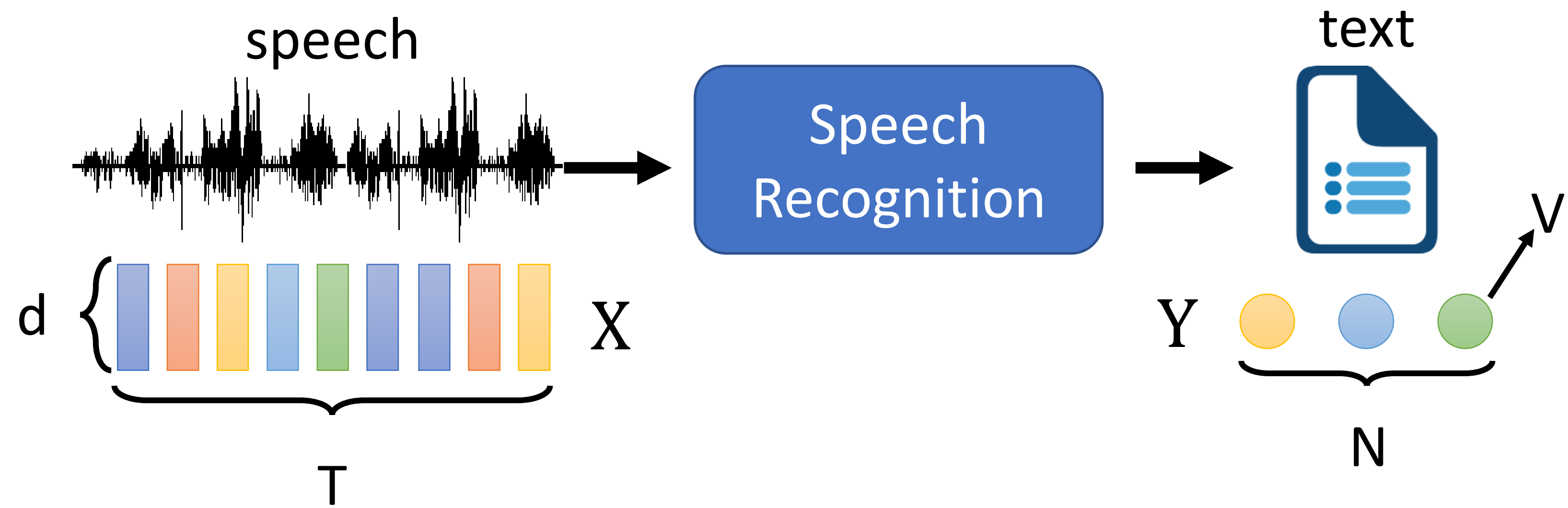


Seq-to-seq



HMM

Hidden Markov Model (HMM)



$$P_\theta(X|\underline{Y}) = ?$$

The token here is small unit called state.

Training

$$\theta^* = arg \max_\theta log P_\theta(X|\hat{Y})$$

Testing

$$Y^* = arg \max_Y log P_\theta(X|Y)$$

HMM

- A sentence corresponds to a sequence of **states**

what do you think

Phoneme:

hh w aa t d uw y uw th ih ng k

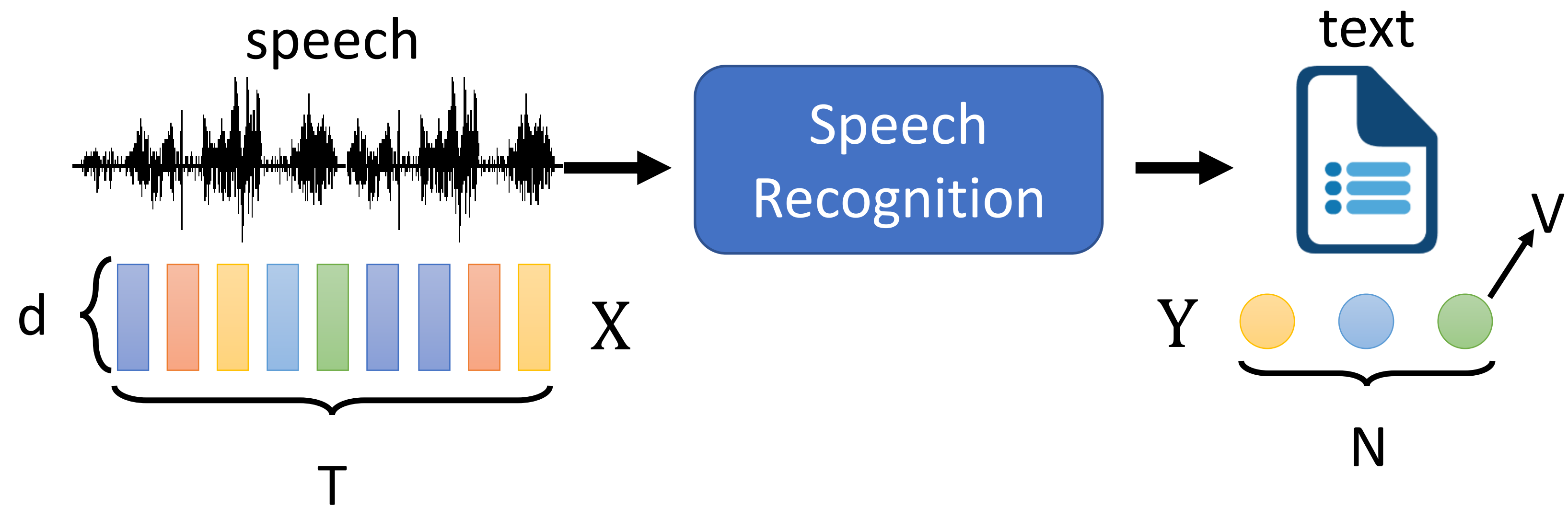
Tri-phone:

…… t-d+uw d-uw+y uw-y+uw y-uw+th ……

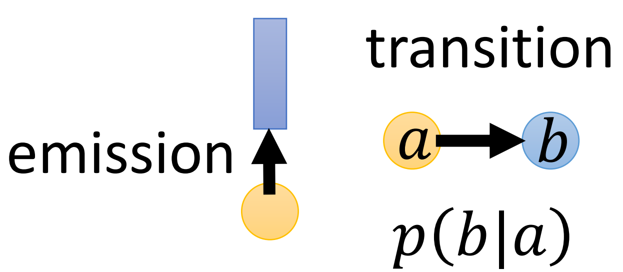
t-d+uw1 t-d+uw2 t-d+uw3 d-uw+y1 d-uw+y2 d-uw+y3

State:

Hidden Markov Model (HMM)



$$P_{\boxed{\theta}}(X|Y) = ?$$



Training

$$\theta^* = arg \max_{\theta} log P_{\theta}(X|\hat{Y})$$

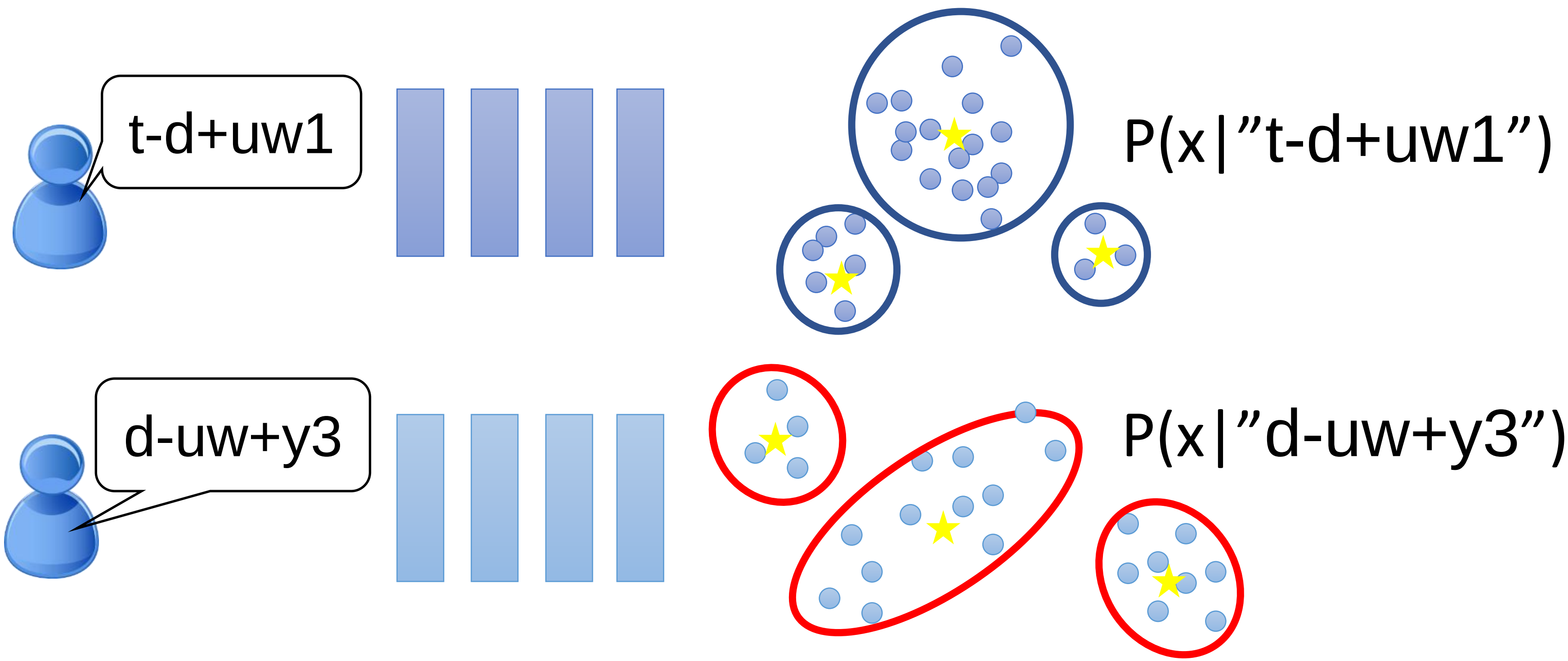
Testing

$$Y^* = arg \max_{Y} log P_{\theta}(X|Y)$$

HMM – Emission Probability

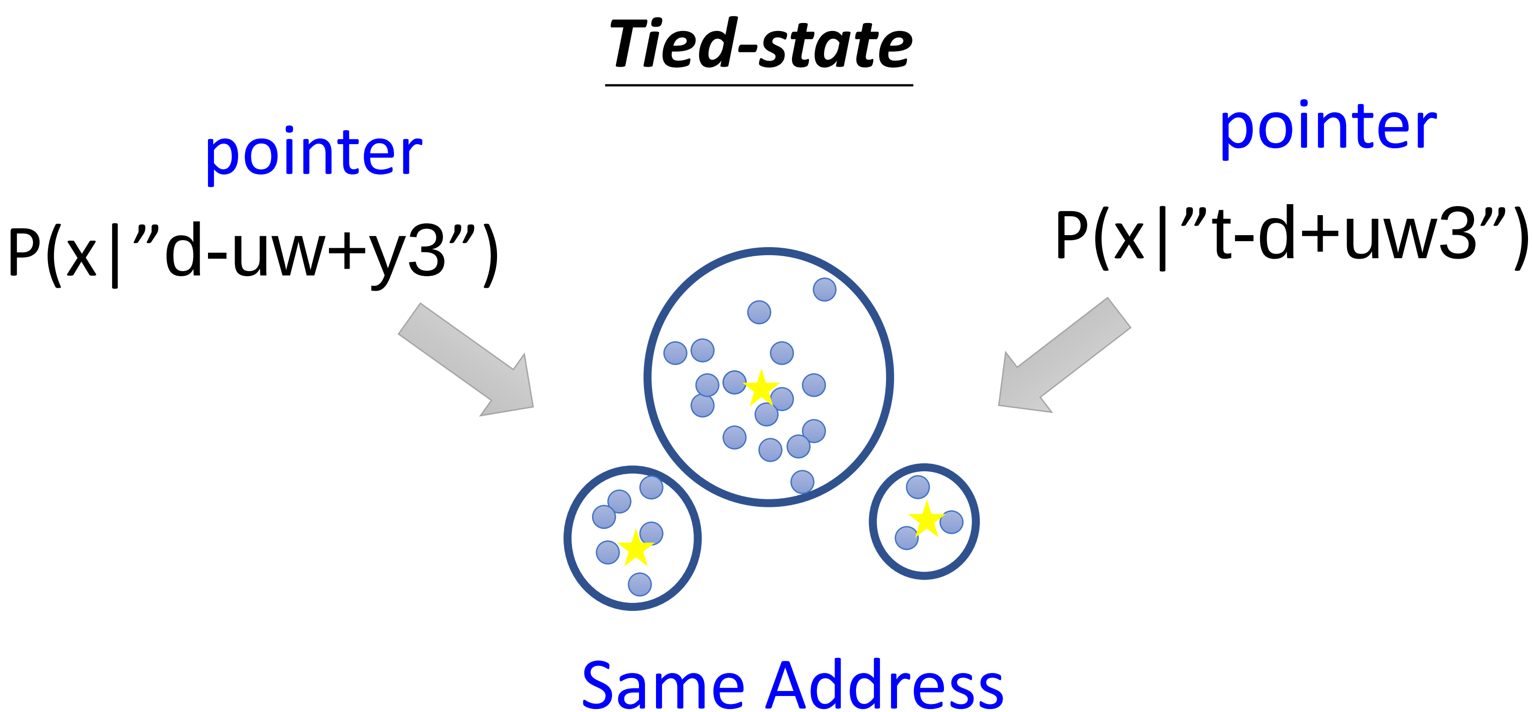
- Each state has a stationary distribution for acoustic features

Gaussian Mixture Model (GMM)



HMM – Emission Probability

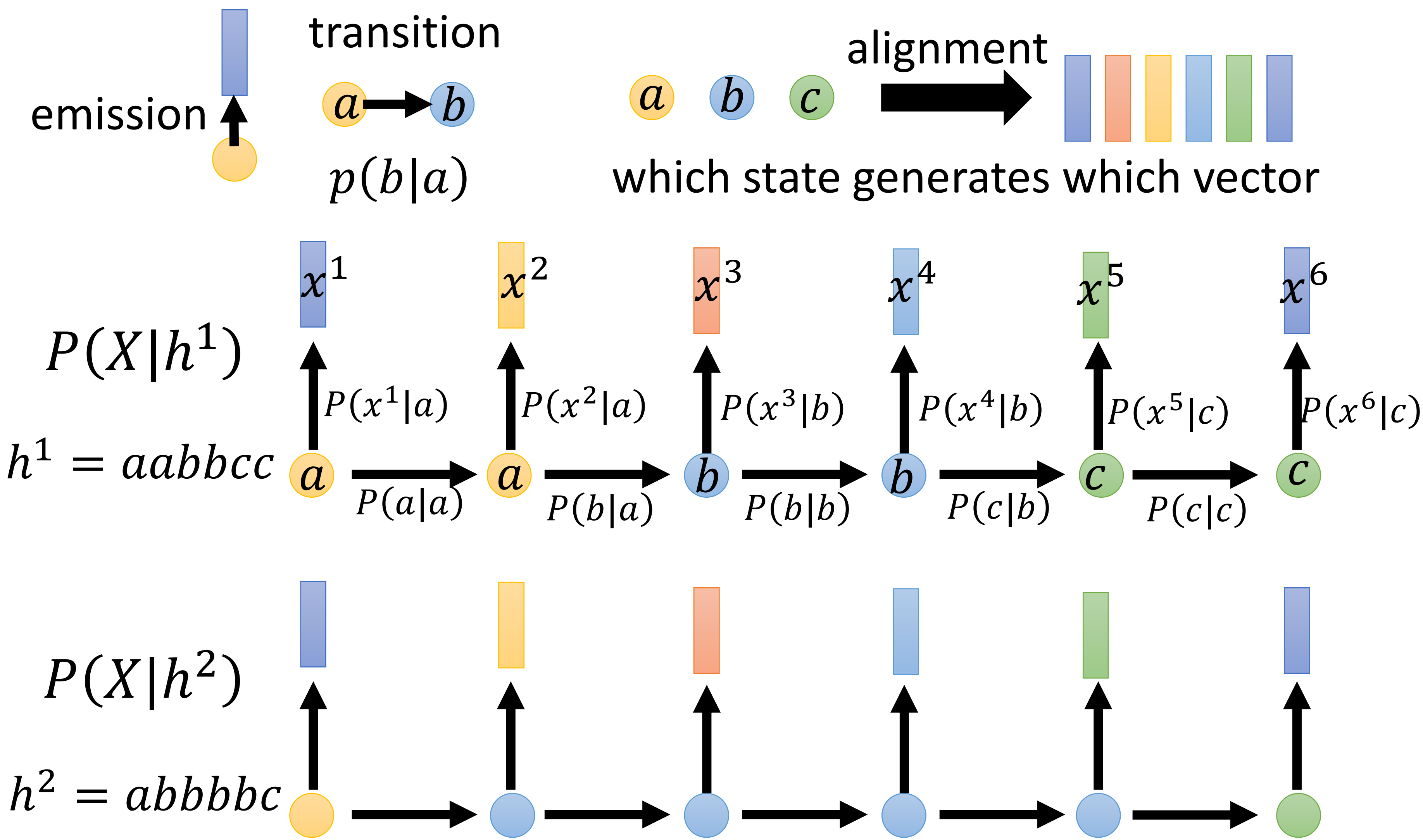
- Each state has a stationary distribution for acoustic features



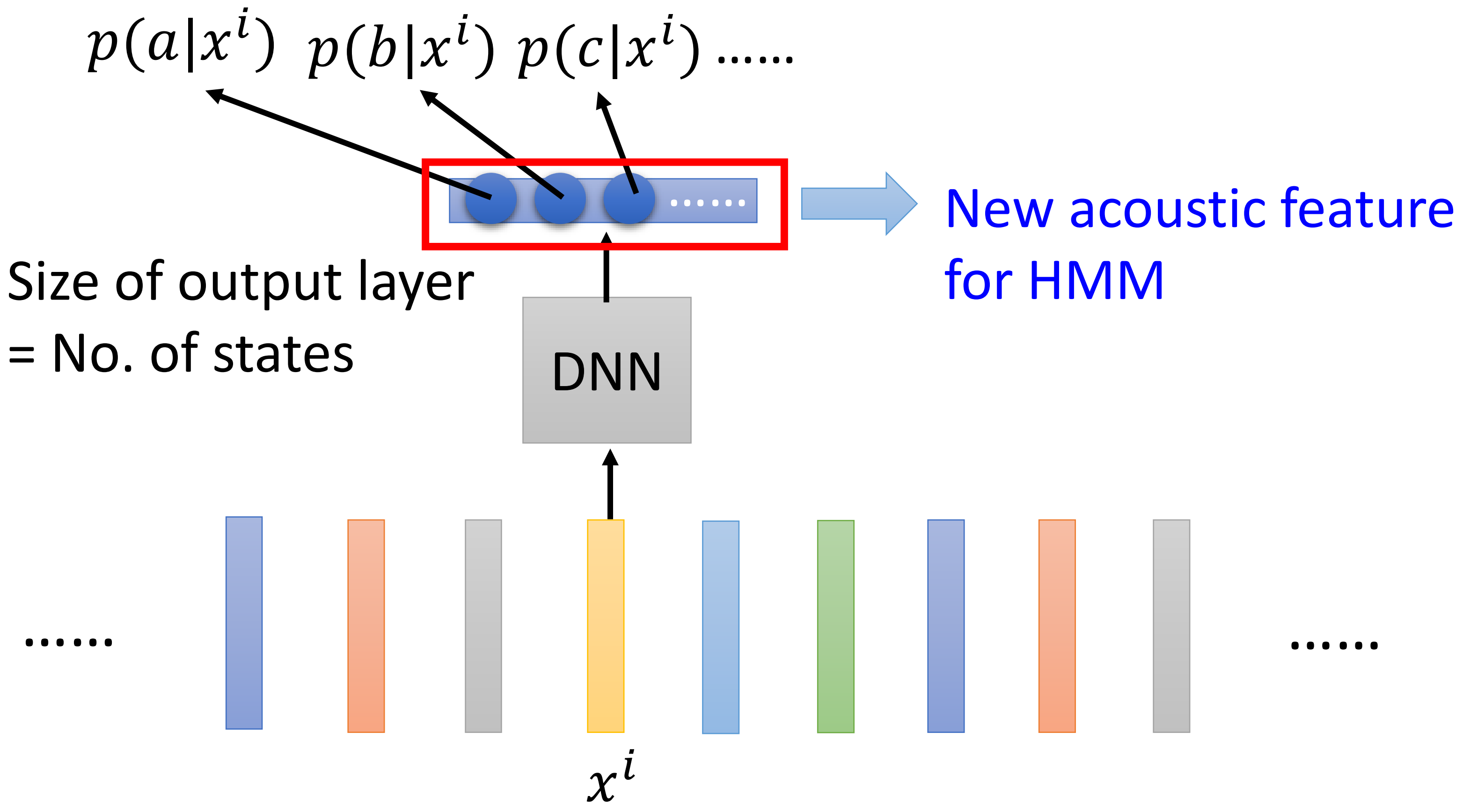
$$P_{\boxed{\theta}}(X|Y) = ? \sum_{h \in align(Y)} P(X|h)$$

$h = abccbc$

$h = abbbcccc$

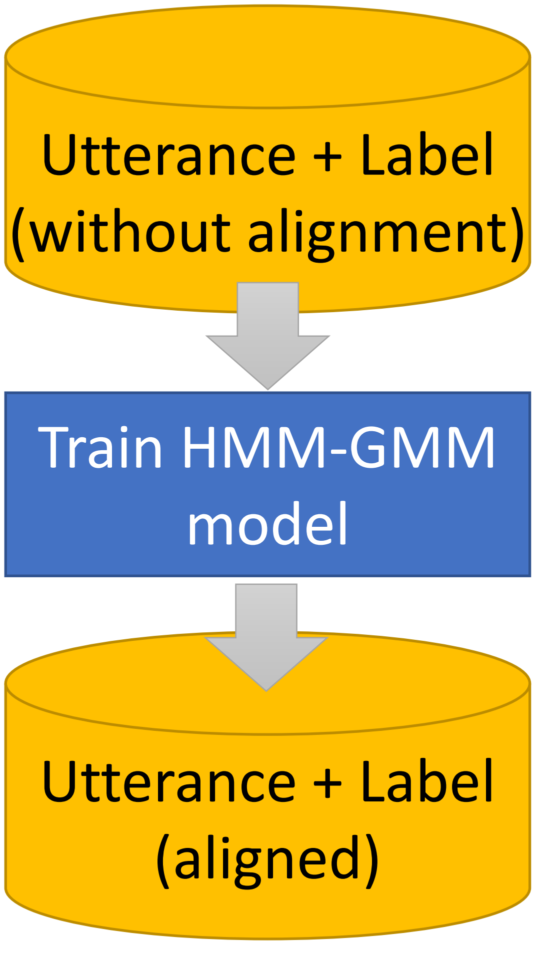


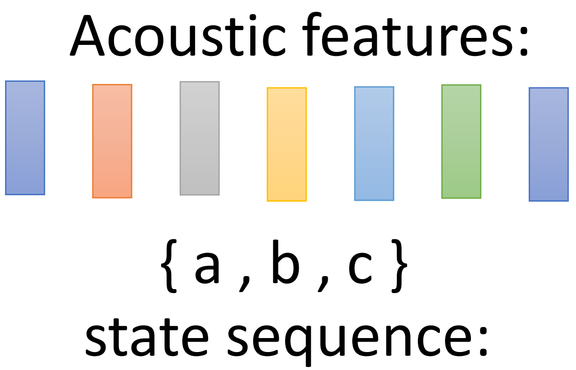
Before End-to-end – Tandem

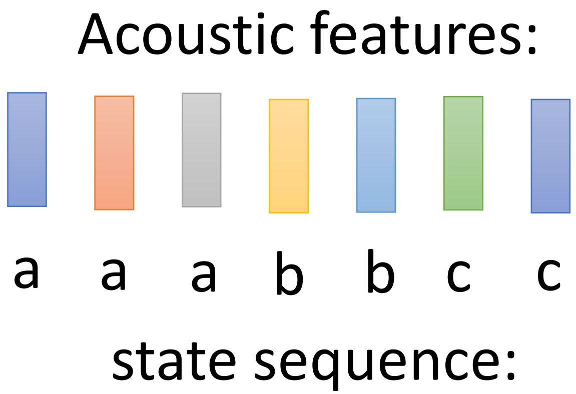


Last hidden layer or bottleneck layer are also possible.

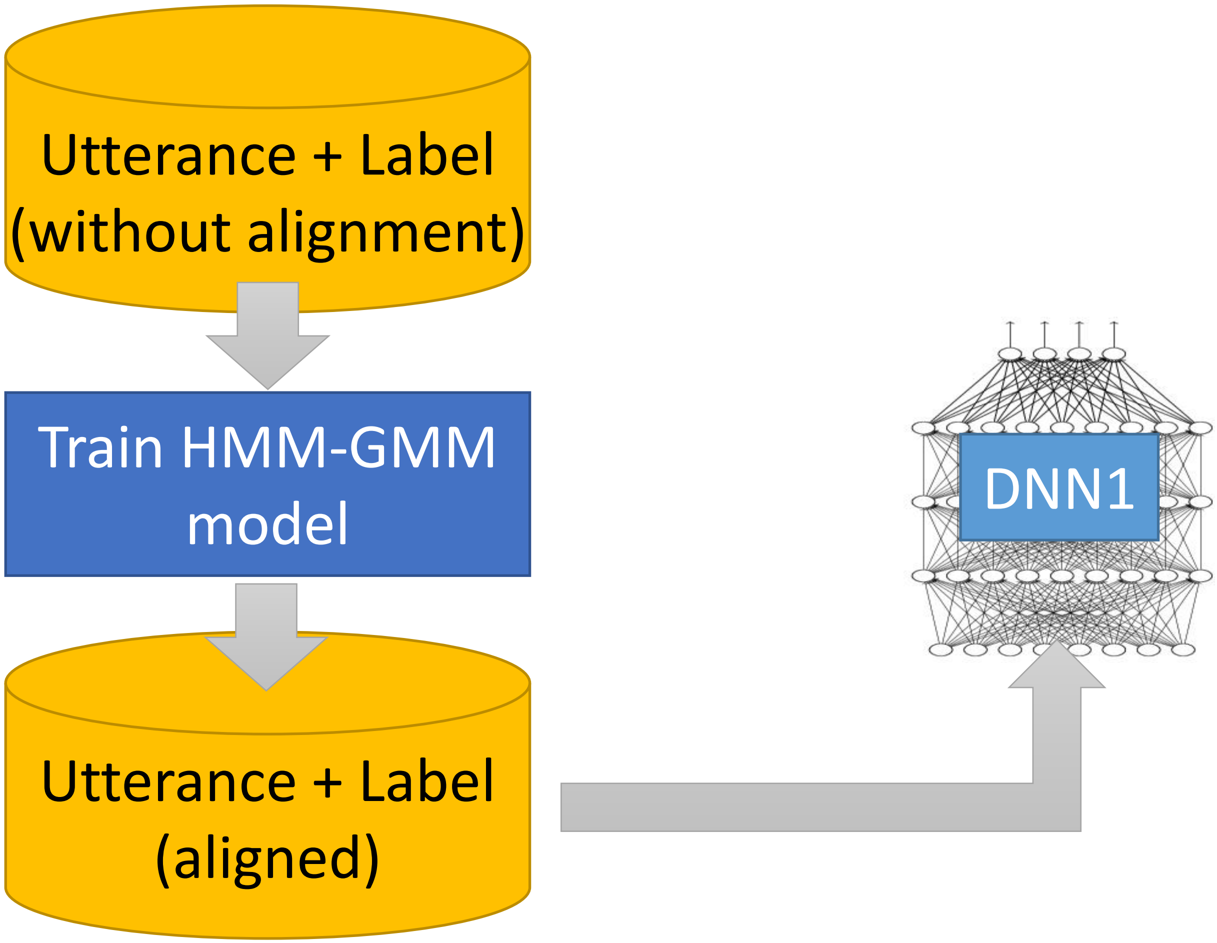
Before End-to-end – Tandem



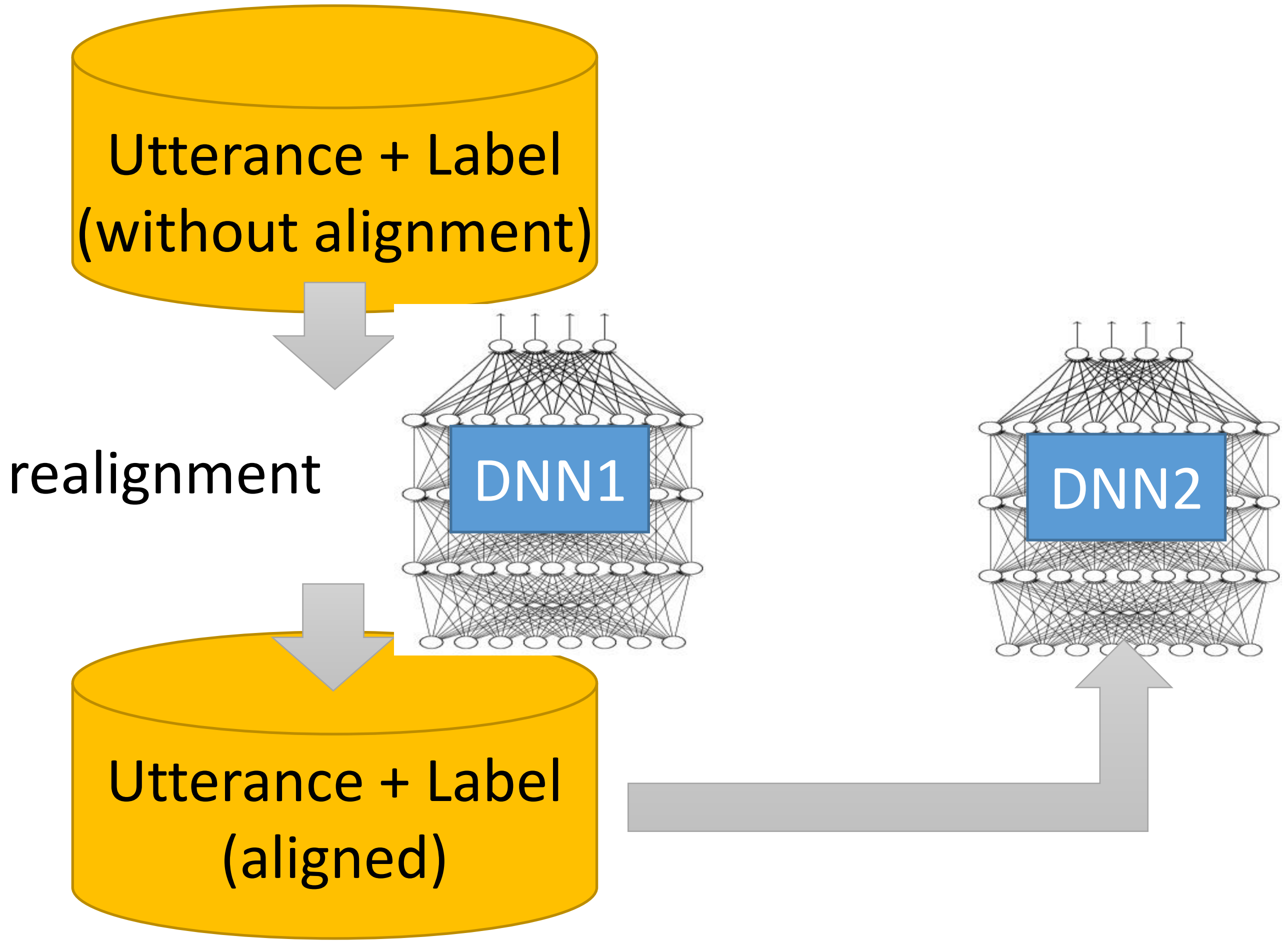




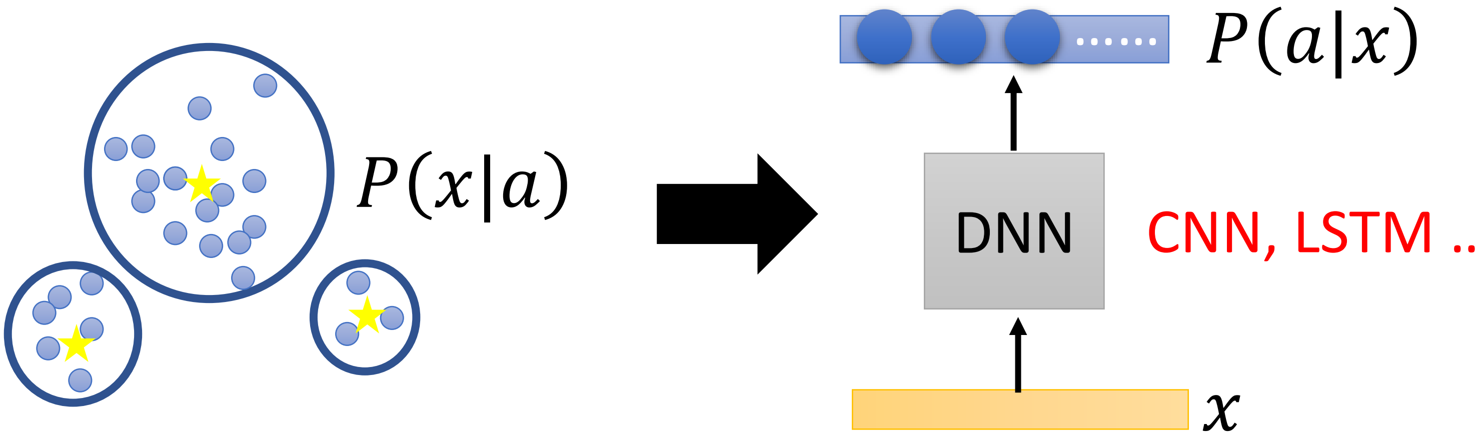
Before End-to-end – Tandem



Before End-to-end – Tandem



Before End-to-end – Hybrid



DNN output

$$P(x|a) = \frac{P(x,a)}{P(a)} = \frac{P(a|x)\cancel{P(x)}}{P(a)}$$

Count from training data

Human Parity!

- 微軟語音辨識技術突破重大里程碑：對話辨識能力達人類水準！(2016.10)
  - https://www.bnext.com.tw/article/41414/bn-2016-10-19-020437-216

Machine 5.9% v.s. Human 5.9%

[Yu, et al., INTERSPEECH'16]

- IBM vs Microsoft: 'Human parity' speech recognition record changes hands again (2017.03)
  - http://www.zdnet.com/article/ibm-vs-microsoft-human-parity-speech-recognition-record-changes-hands-again/

Machine 5.5% v.s. Human 5.1%

[Saon, et al., INTERSPEECH'17]

Very Deep

VGG Net (85M Parameters)	Residual-Net (38M Parameters)	LACE (65M Parameters)
14 weight layers	49 weight layers	22 weight layers
40x41 input	40x41 input	40x61 input
3 – conv 3x3, 96	3 – [conv 1x1, 64 conv 3x3, 64 conv 1x1, 256]	5 – conv 3x3, 128
Max pool	4 – [conv 1x1, 128 conv 3x3, 128 conv 1x1, 512]	5 – conv 3x3, 256
4 – conv 3x3, 192	6 – [conv 1x1, 256 conv 3x3, 256 conv 1x1, 1024]	5 – conv 3x3, 512
Max pool	3 – [conv 1x1, 512 conv 3x3, 512 conv 1x1, 2048]	5 – conv 3x3, 1024
4 – conv 3x3, 384	Average pool	1 – conv 3x4, 1
Max pool	Softmax (9000)	Softmax (9000)
2 – FC – 4096		
Softmax (9000)		

[Yu, et al., INTERSPEECH'16]

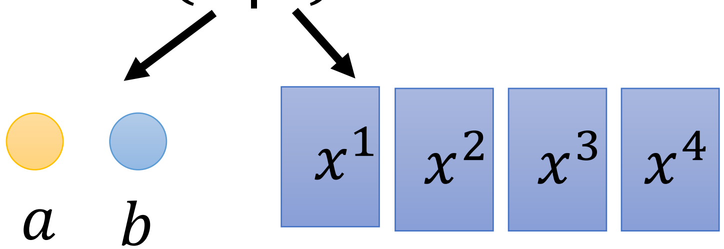
LAS

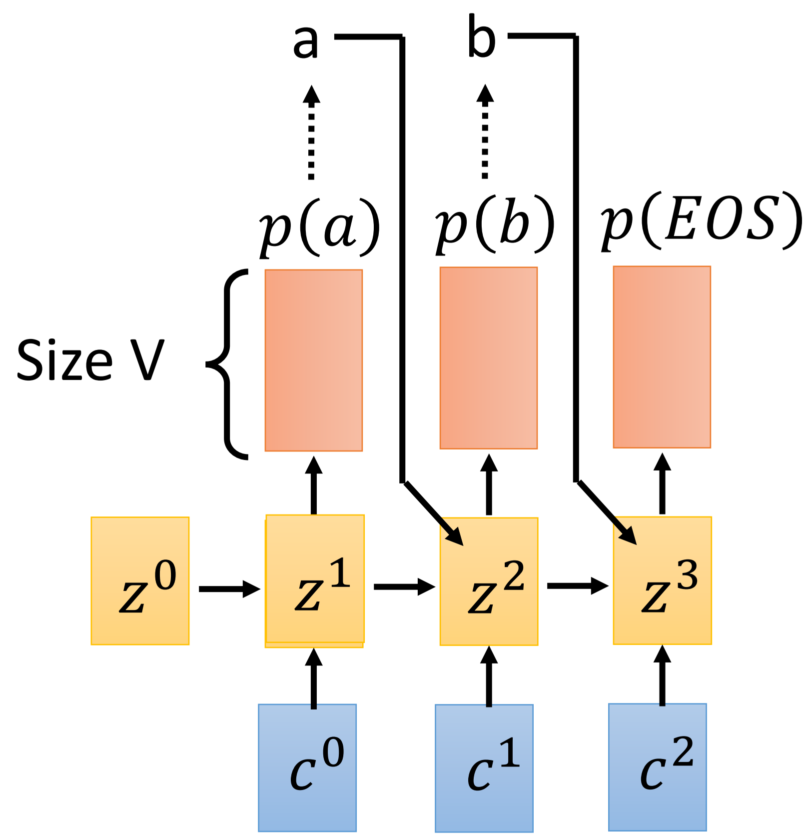
$$\theta^* = arg\max_{\theta} log\mathrm{P}_\theta\left(\hat{Y}|X\right)$$

$$\mathrm{Y}^* = arg\max_{\mathrm{Y}} log\mathrm{P}_\theta(\mathrm{Y}|X)$$

Beam Search

$P(Y|\mathrm{X}) =?$





- LAS directly computes $P(Y|\mathrm{X})$

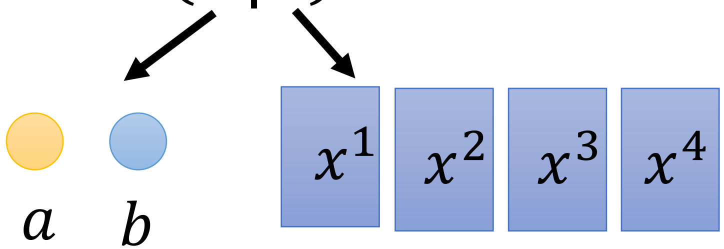
$$P(Y|\mathrm{X}) = p(a|X)p(b|a,X)\ldots$$

CTC, RNN-T

$$\theta^* = arg \max_{\theta} log \mathrm{P}_{\theta}(\hat{Y}|X)$$

$$\mathrm{Y}^* = arg \max_{\mathrm{Y}} log \mathrm{P}_{\theta}(\mathrm{Y}|X)$$

$$P(Y|\mathrm{X}) = ?$$



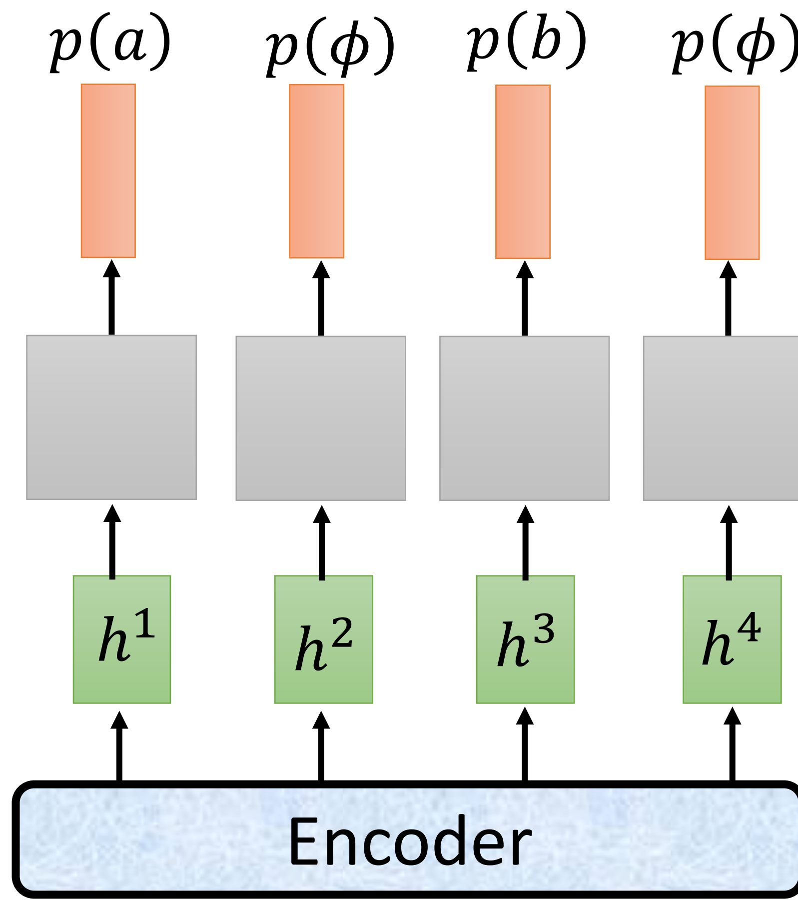
- LAS directly computes $P(Y|\mathrm{X})$

  $$P(Y|\mathrm{X}) = p(a|X)p(b|a,X)\ldots$$

- CTC and RNN-T need ***alignment***

  $$\mathrm{P}_{\theta}(\mathrm{Y}|X) = \sum_{h \in align(Y)} P(h|X)$$

$P(h|X)$ $\quad h = a\ \phi\ b\ \phi \rightarrow a\ b$



HMM, CTC, RNN-T

HMM

$$\mathrm{P}_\theta(X|\mathrm{Y}) = \boxed{\sum_{\underline{h \in align(Y)}} P(X|h)}$$

CTC, RNN-T

$$\mathrm{P}_\theta(\mathrm{Y}|X) = \boxed{\sum_{\underline{h \in align(Y)}} P(h|X)}$$

1. Enumerate all the possible alignments

2. How to sum over all the alignments

3. Training: $\boxed{\theta^* = arg \max_\theta log \mathrm{P}_\theta\left(\hat{Y}|X\right)}$ $\quad \dfrac{\partial P\left(\hat{Y}|X\right)}{\partial \theta} = ?$

4. Testing (Inference, decoding):

$$\mathrm{Y}^* = arg \max_{\mathrm{Y}} log \mathrm{P}_\theta(\mathrm{Y}|X)$$

All the alignments

LAS

你們在忙什麼 ☺

	Y		$h \in align(Y)$
HMM	c a t	duplicate to length T	c c a a a t　c a a a a t　…
CTC	c a t	duplicate / add ϕ to length T	c ϕ a a t t　ϕ c a ϕ t ϕ　…
RNN-T	c a t	add ϕ x T	c ϕ ϕ ϕ a ϕ ϕ t ϕ c ϕ ϕ a ϕ ϕ t ϕ ϕ ……

HMM c a t → c c a a a t c a a a a t …

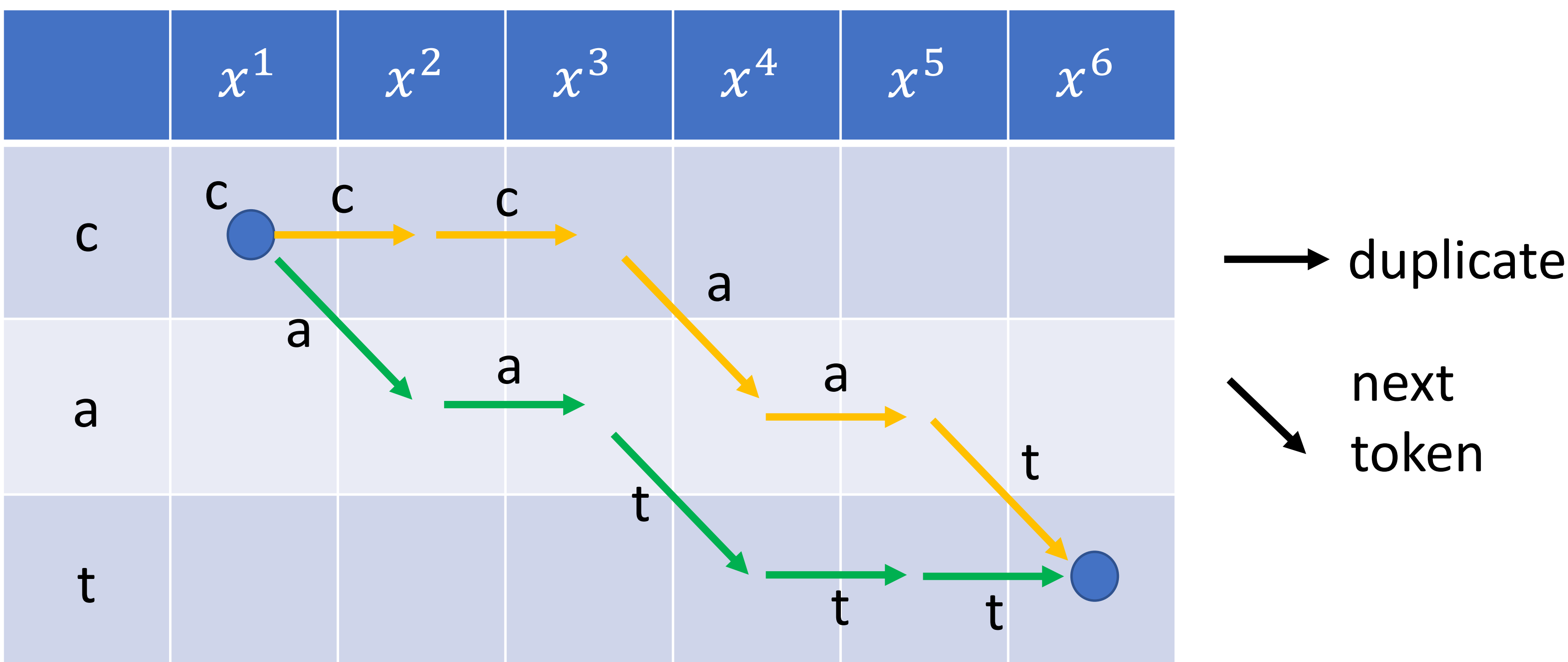
duplicate to length T

For n = 1 to N

  output the n-th token t_n times

constraint: $t_1 + t_2 + \cdots t_N = T$, $t_n > 0$

Trellis Graph



HMM c a t → c c a a a t   c a a a a t   …

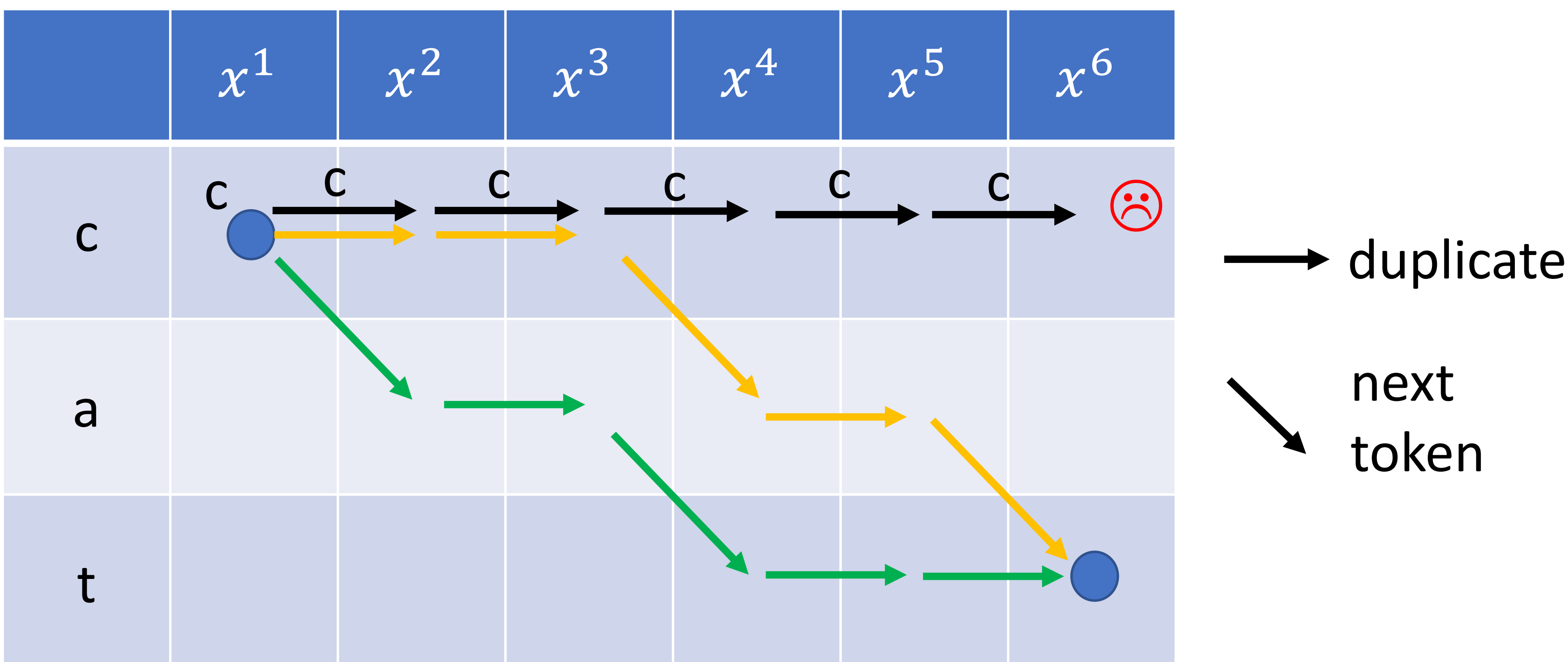
duplicate to length T

For n = 1 to N

output the n-th token t_n times

constraint: $t_1 + t_2 + \cdots t_N = T,\ t_n > 0$

Trellis Graph



CTC c a t $\xrightarrow{\hspace{4em}}$ c ϕ a a t t ϕ c a ϕ t ϕ …

$\left\{\begin{array}{l}\text{duplicate}\\ \text{add } \phi\end{array}\right.$ to length T

output "ϕ" c_0 times

For n = 1 to N

output the n-th token t_n times

output "ϕ" c_n times

constraint: $t_1 + t_2 + \cdots t_N +$

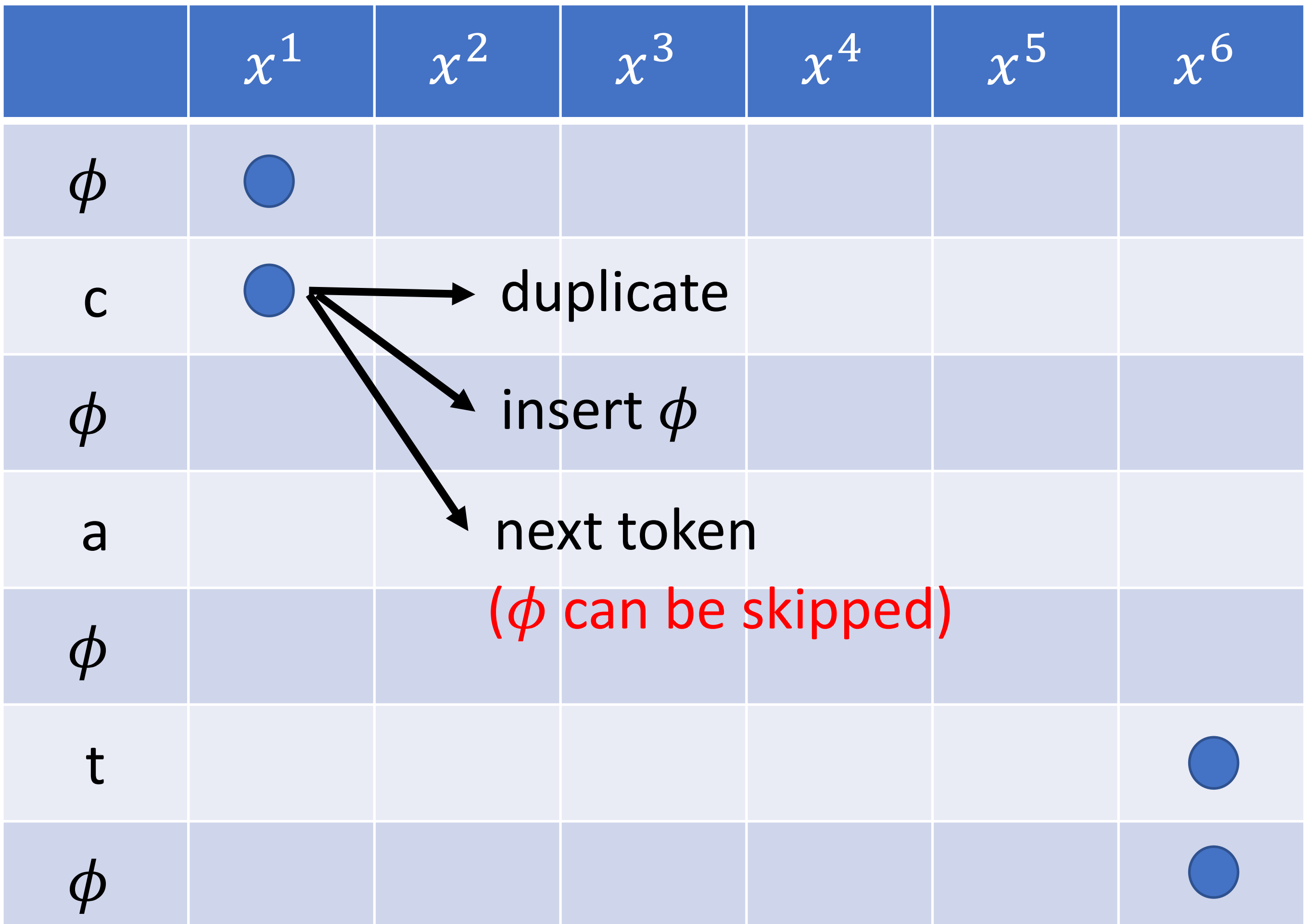
$c_0 + c_1 + \cdots c_N = T$

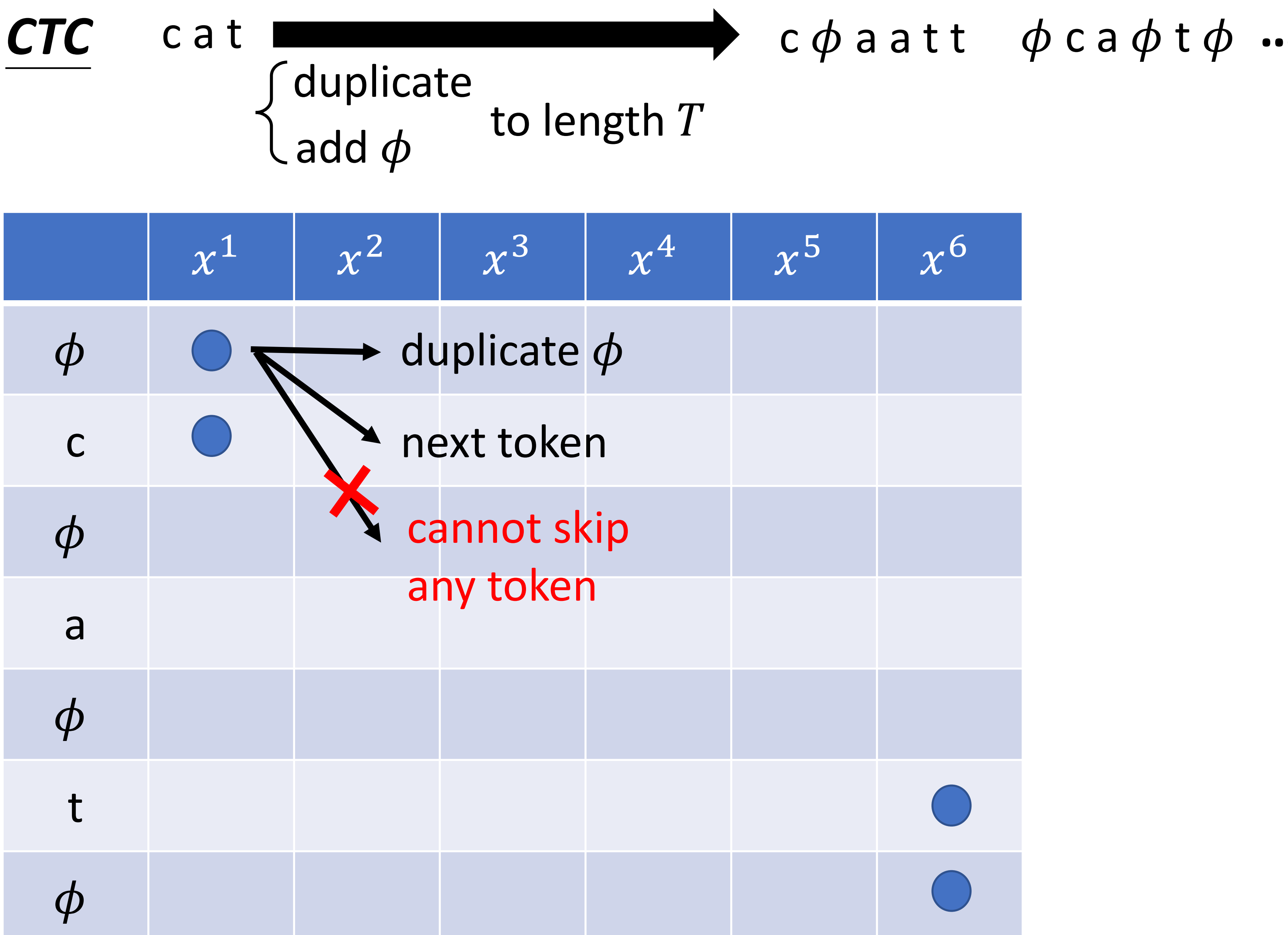
$t_n > 0 \quad c_n \geq 0$

CTC

c a t $\longrightarrow$ c ϕ a a t t ϕ c a ϕ t ϕ …

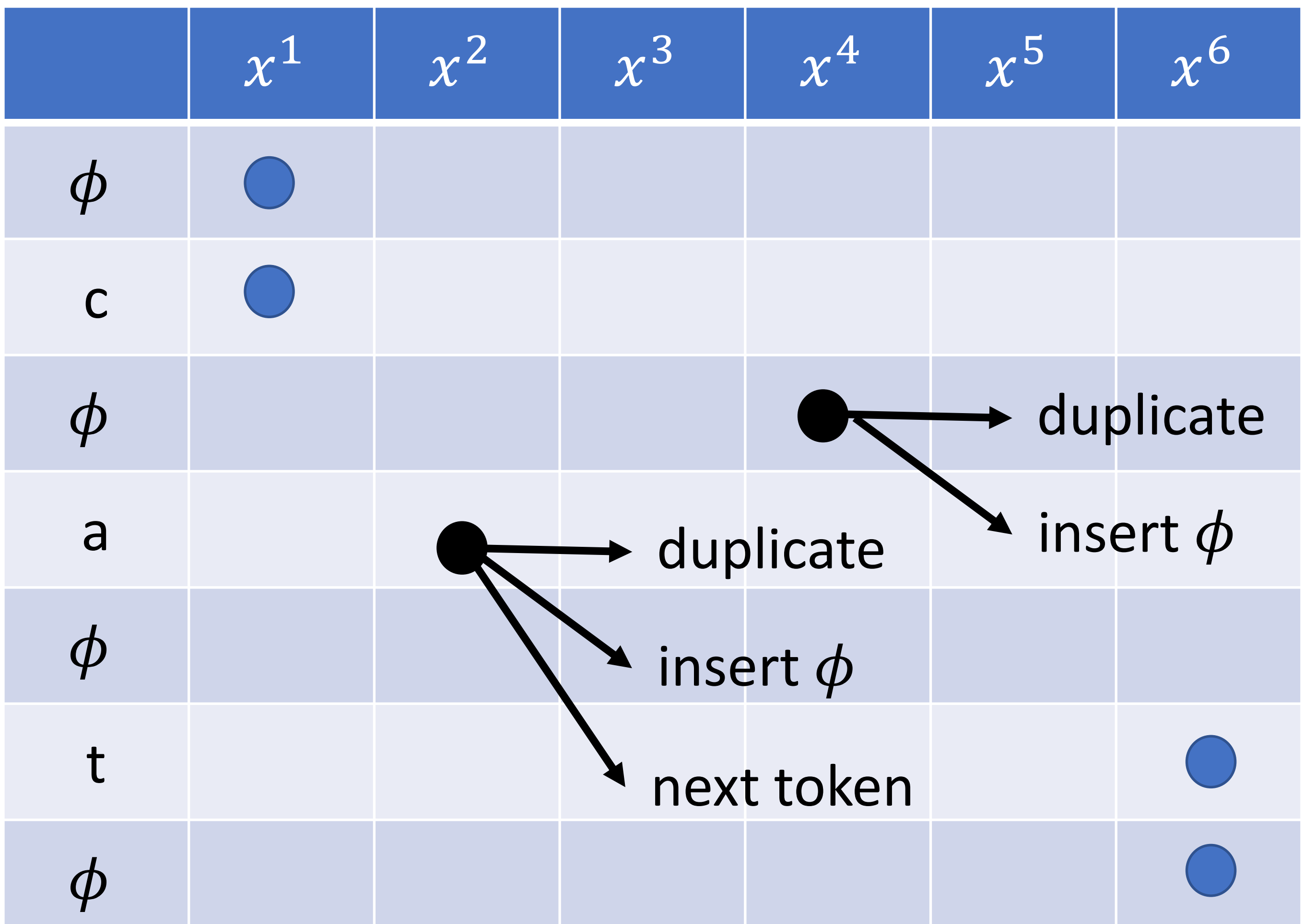
$\left\{\begin{array}{l}\text{duplicate}\\ \text{add } \phi\end{array}\right.$ to length T




CTC
c a t
c ϕ a a t t
ϕ c a ϕ t ϕ …
duplicate
add ϕ
to length T
x1 x2 x3 x4 x5 x6
ϕ
c
ϕ
a
ϕ
t
ϕ
duplicate ϕ
next token
cannot skip
any token

CTC

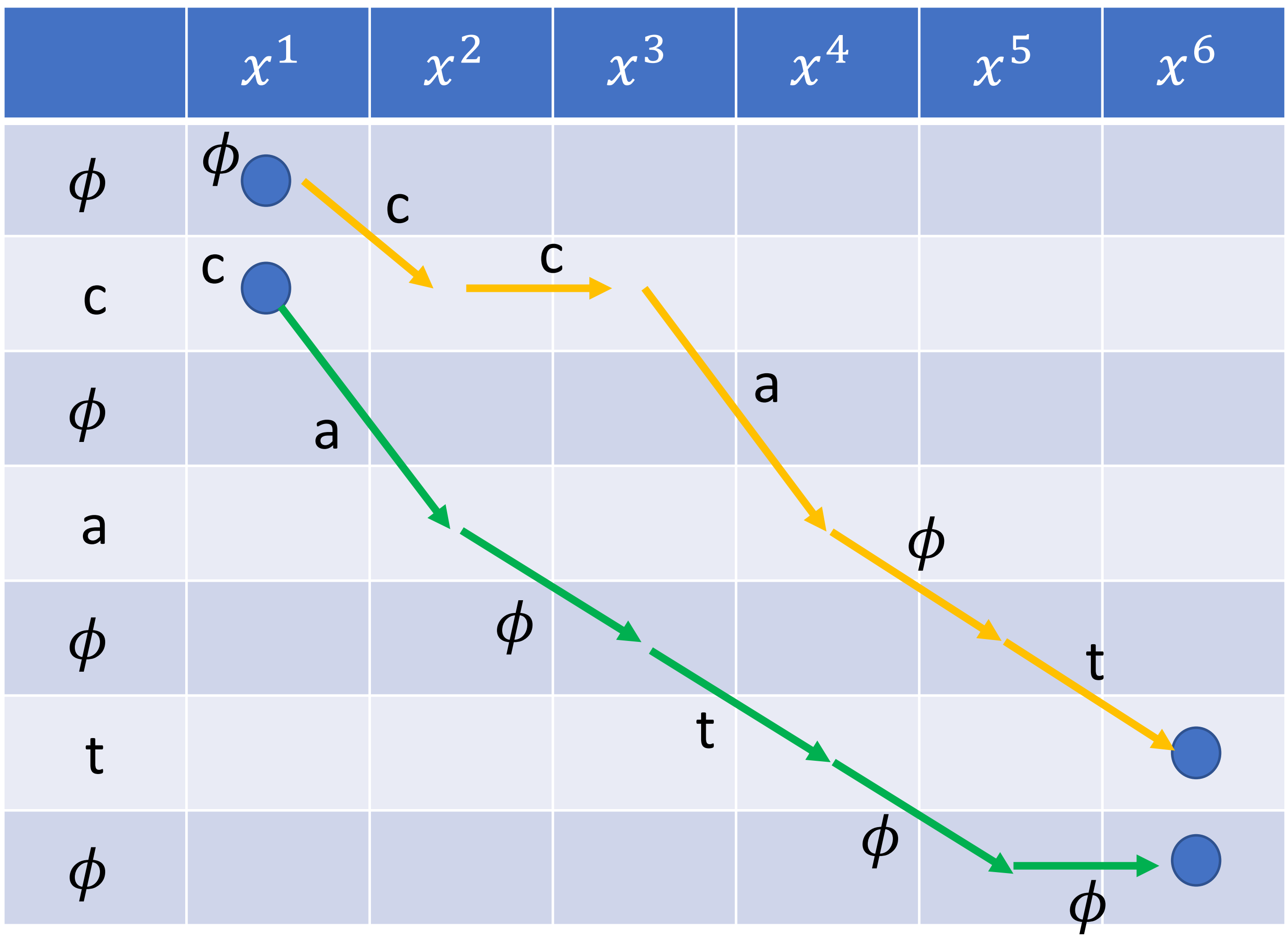
c a t $\xrightarrow{\left\{\begin{array}{l}\text{duplicate} \\ \text{add } \phi\end{array}\right. \text{ to length } T}$ c ϕ a a t t ϕ c a ϕ t ϕ …



CTC

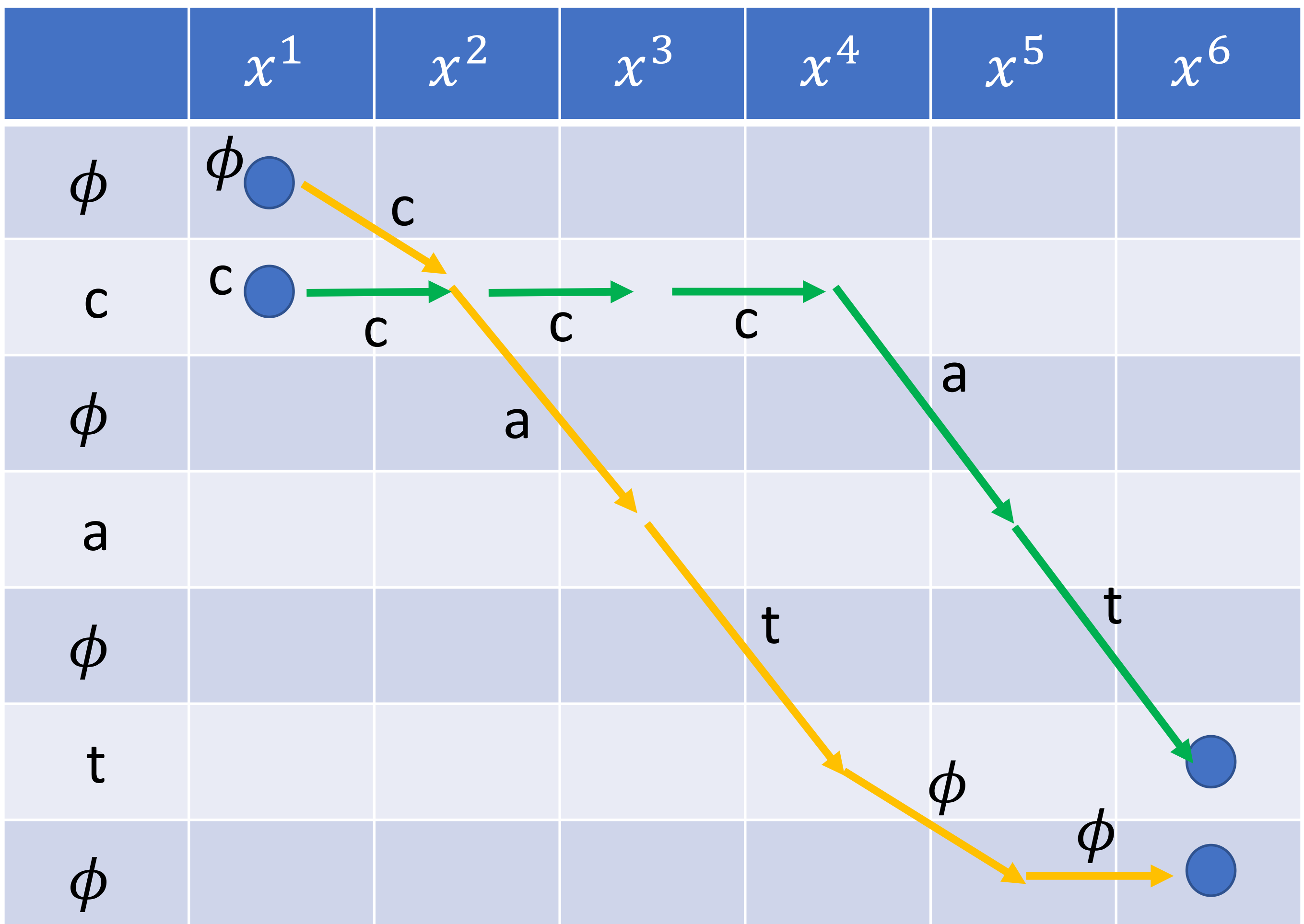
c a t $\longrightarrow$ c ϕ a a t t ϕ c a ϕ t ϕ ...

$\left\{\begin{array}{l}\text{duplicate}\\ \text{add } \phi\end{array}\right.$ to length T



CTC c a t → c ϕ a a t t ϕ c a ϕ t ϕ …

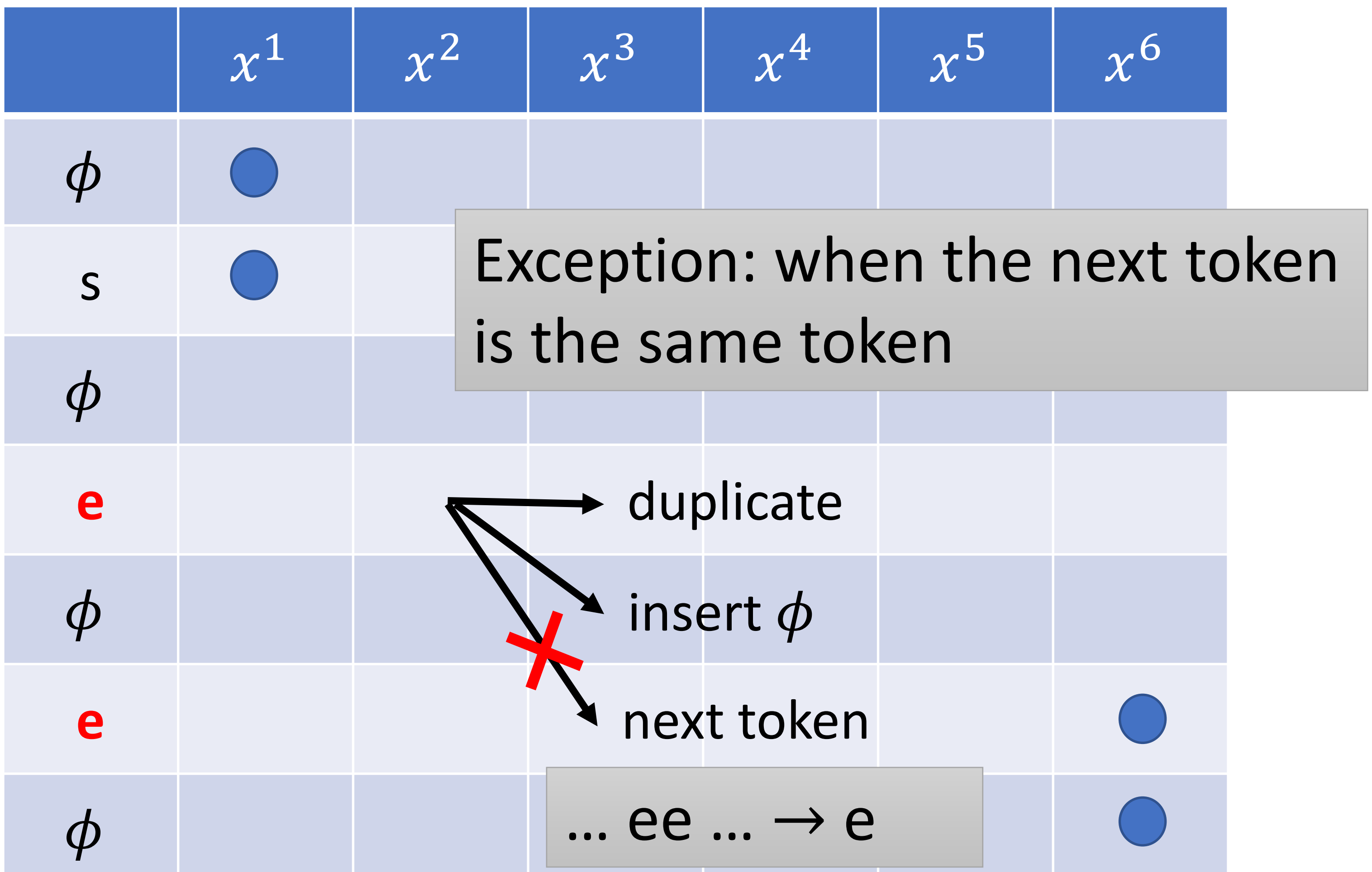
$\left\{\begin{array}{l}\text{duplicate}\\ \text{add } \phi\end{array}\right.$ to length T



CTC

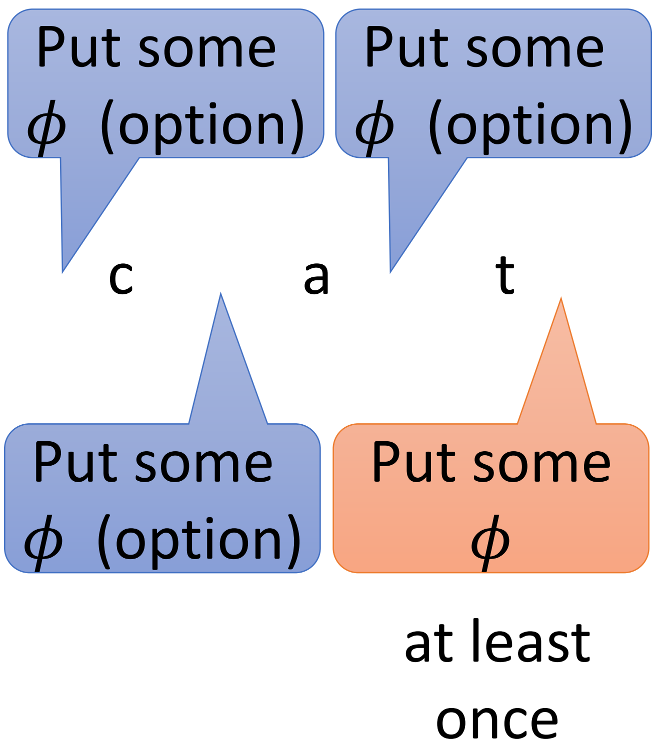
c a t $\longrightarrow$ c ϕ a a t t ϕ c a ϕ t ϕ …

$\left\{\begin{array}{l}\text{duplicate}\\ \text{add } \phi\end{array}\right.$ to length T



RNN-T c a t → c ϕ ϕ ϕ a ϕ ϕ t ϕ ……

add ϕ x T → c ϕ ϕ a ϕ ϕ t ϕ ϕ



output "ϕ" c_0 times

For n = 1 to N

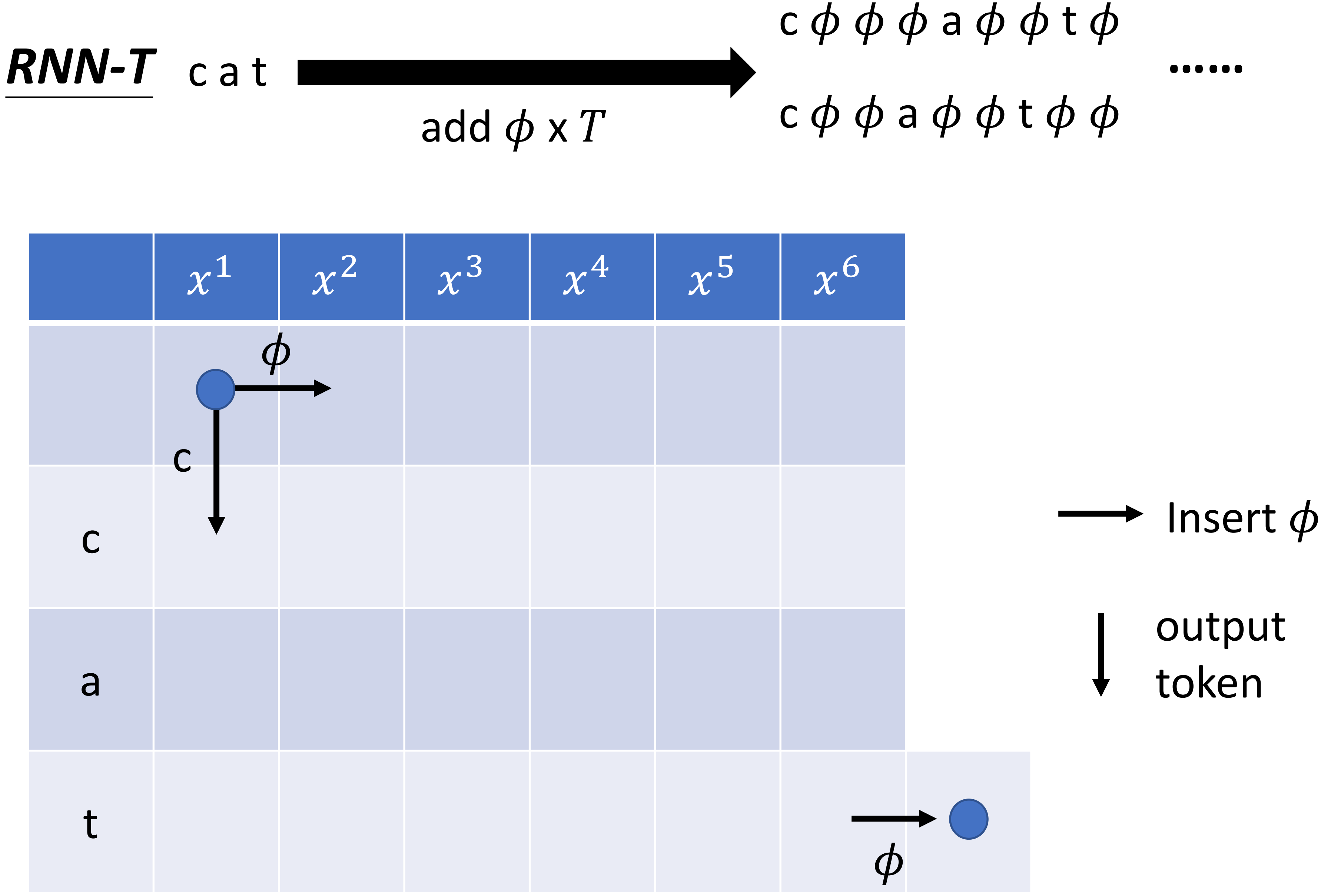
output the n-th token 1 times

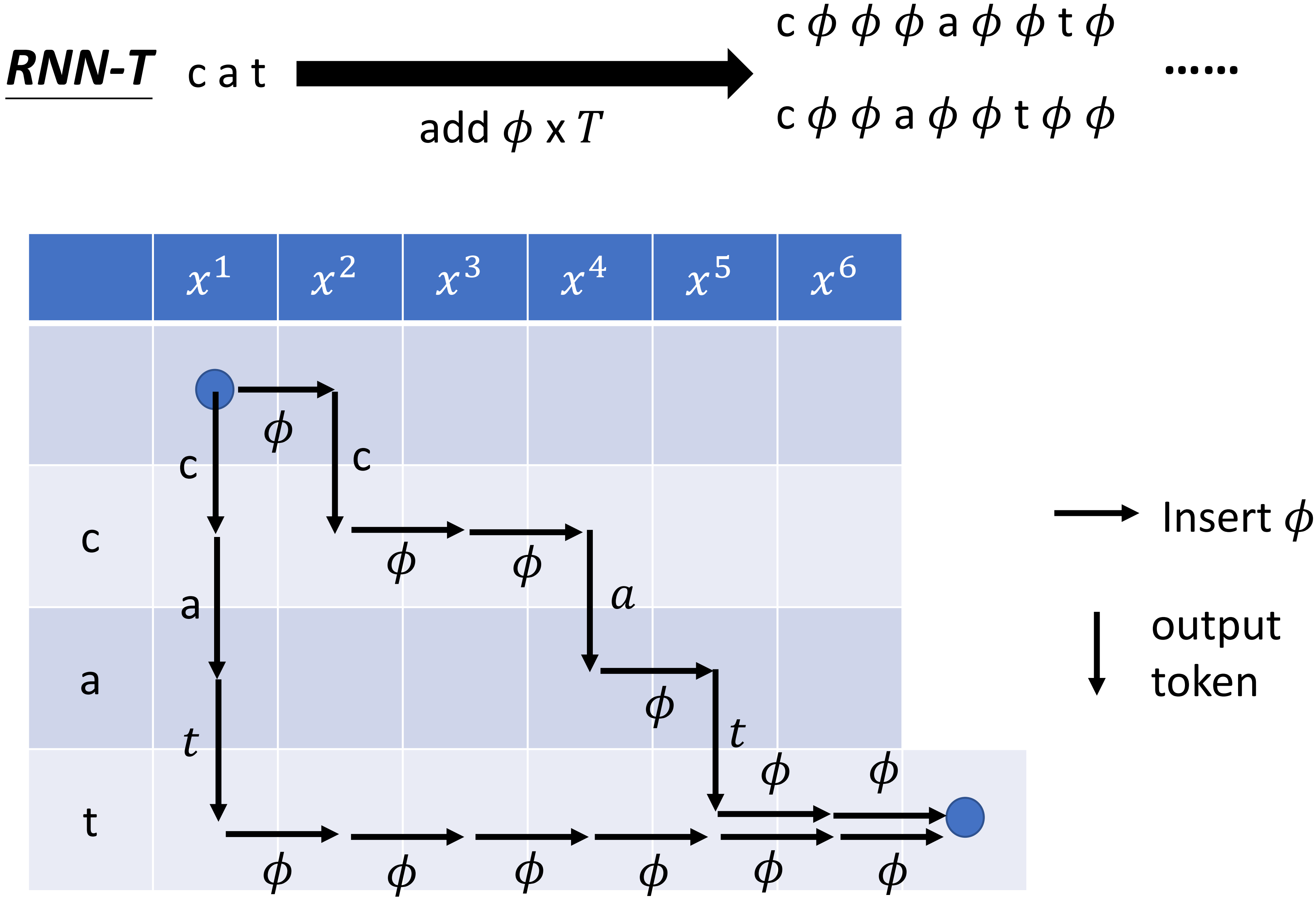
output "ϕ" c_n times

constraint: $c_0 + c_1 + \cdots c_N = T$

$c_N > 0$

$c_n \geq 0$ for n = 1 to $N - 1$


RNN-T
c a t
add ϕ x T
c ϕ ϕ ϕ a ϕ ϕ t ϕ
c ϕ ϕ a ϕ ϕ t ϕ ϕ
……
x^1
x^2
x^3
x^4
x^5
x^6
c
a
t
ϕ
c
ϕ
Insert ϕ
output token

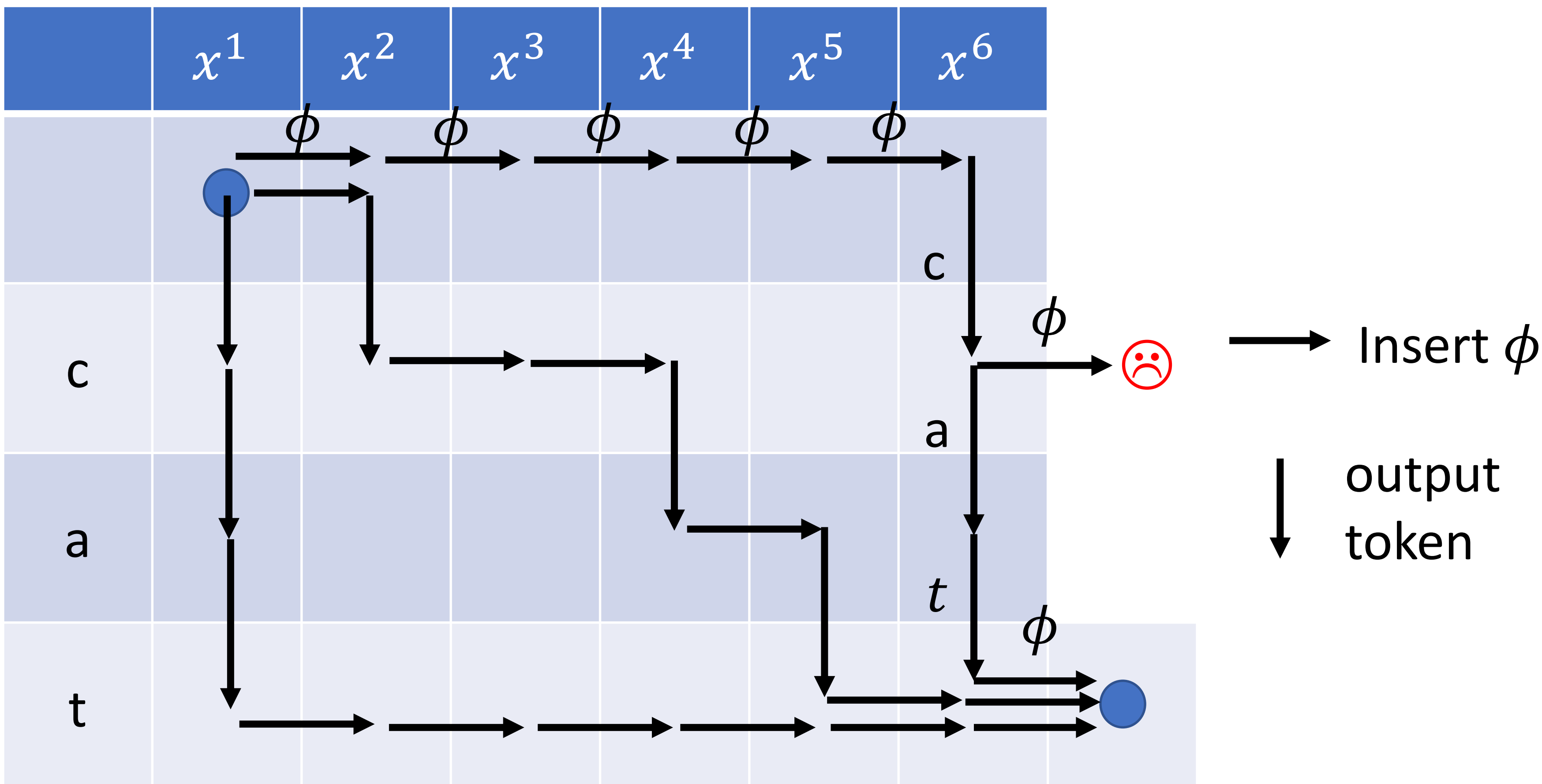

RNN-T
c a t
add ϕ x T
c ϕ ϕ ϕ a ϕ ϕ t ϕ
c ϕ ϕ a ϕ ϕ t ϕ ϕ
……
x1 x2 x3 x4 x5 x6
c
a
t
ϕ
c
a
t
Insert ϕ
output token

RNN-T c a t $\xrightarrow[\text{add } \phi \text{ x } T]{}$

c ϕ ϕ ϕ a ϕ ϕ t ϕ

c ϕ ϕ a ϕ ϕ t ϕ ϕ

......



HMM, CTC, RNN-T

HMM

$$\mathrm{P}_\theta(X|\mathrm{Y}) = \boxed{\sum_{\underline{h \in align(Y)}} P(X|h)}$$

CTC, RNN-T

$$\mathrm{P}_\theta(\mathrm{Y}|X) = \boxed{\sum_{\underline{h \in align(Y)}} P(h|X)}$$

1. Enumerate all the possible alignments

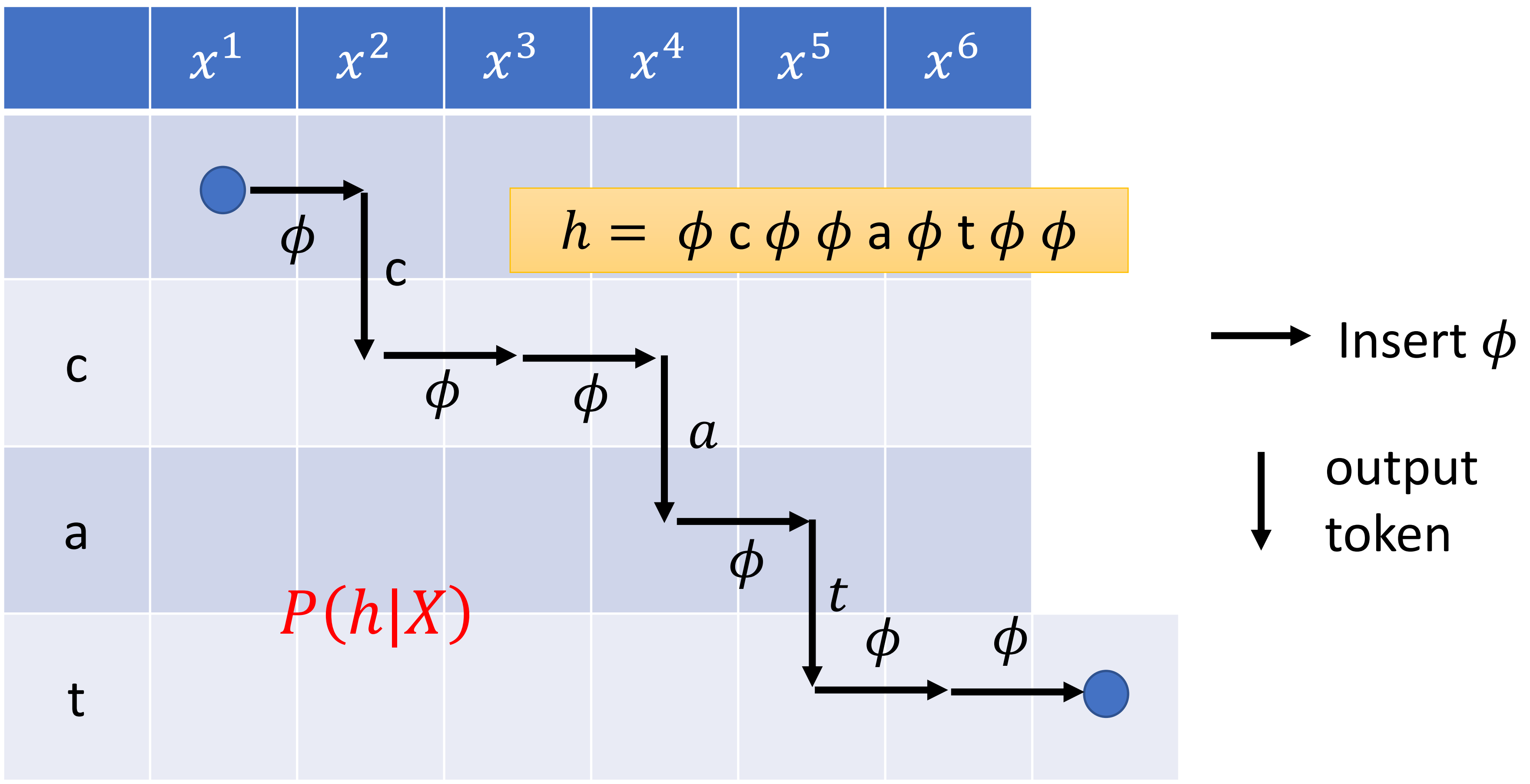
2. How to sum over all the alignments

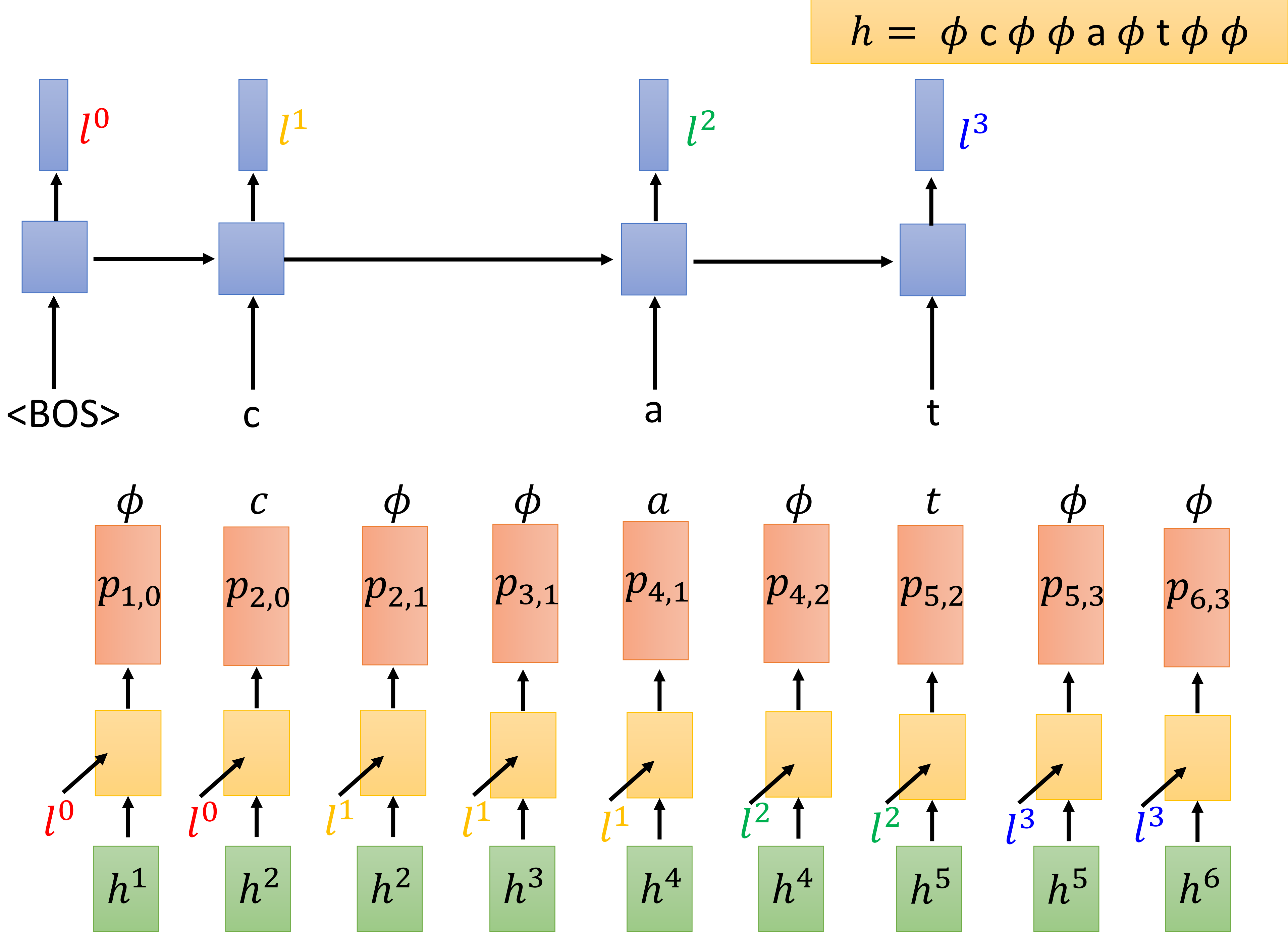
3. Training: $\boxed{\theta^* = arg \max_\theta log \mathrm{P}_\theta\left(\hat{Y}|X\right)}$ $\qquad \frac{\partial P\left(\hat{Y}|X\right)}{\partial \theta} = ?$

4. Testing (Inference, decoding):

$$\mathrm{Y}^* = arg \max_{\mathrm{Y}} log \mathrm{P}_\theta(\mathrm{Y}|X)$$

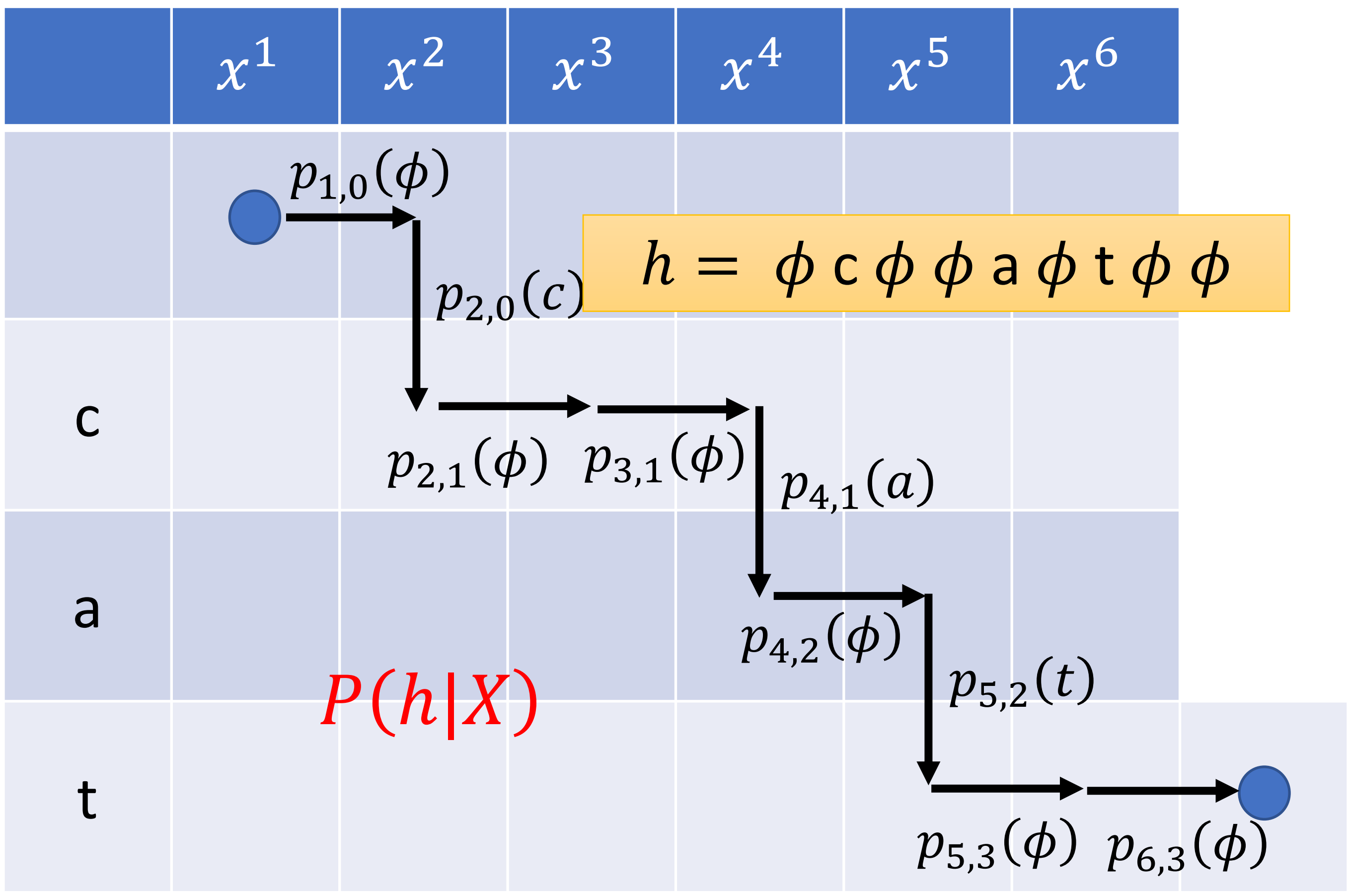
Score Computation



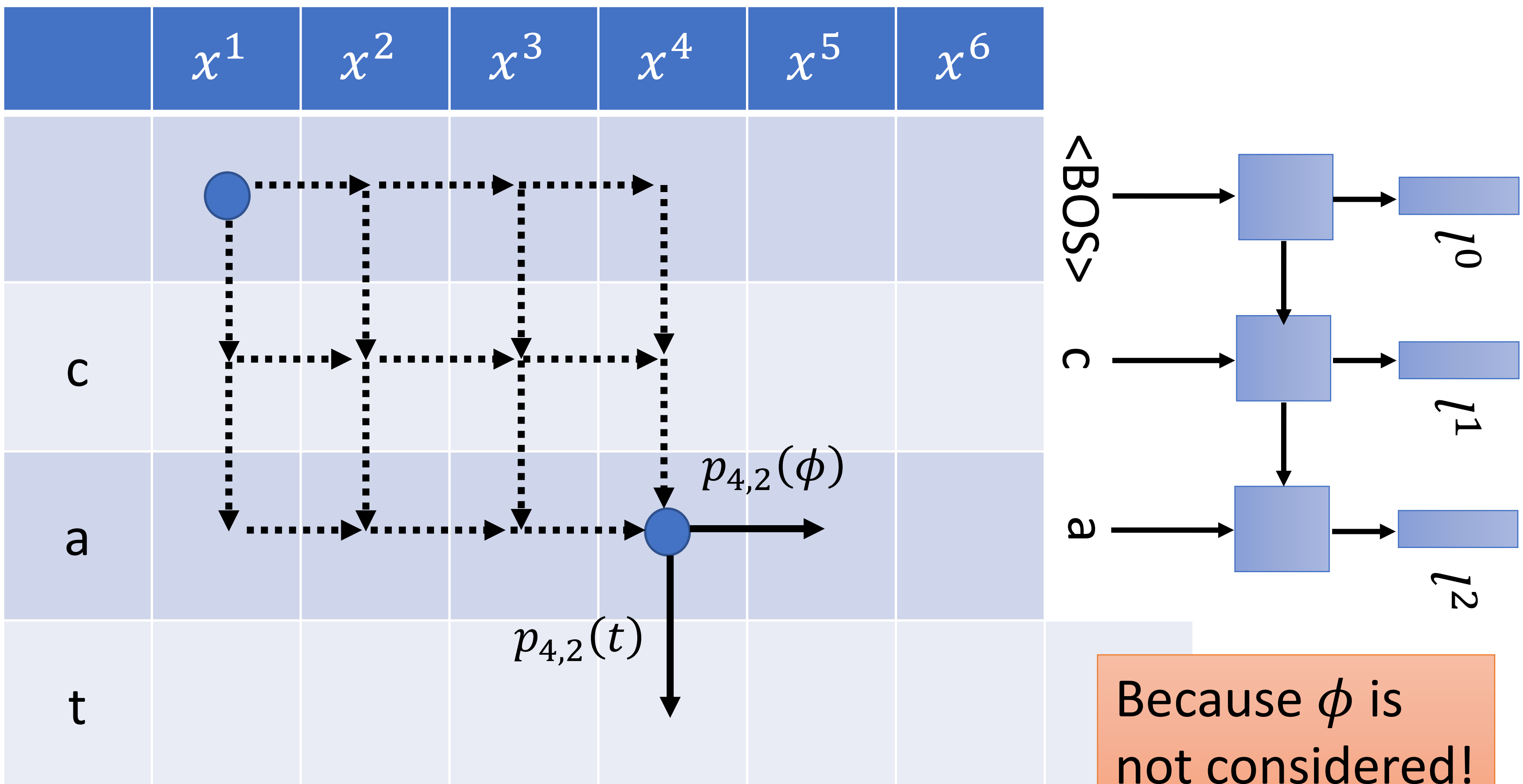


$h = \phi\ c\ \phi\ \phi\ a\ \phi\ t\ \phi\ \phi$
l^0
l^1
l^2
l^3
<BOS>
c
a
t
ϕ c ϕ ϕ a ϕ t ϕ ϕ
$p_{1,0}$ $p_{2,0}$ $p_{2,1}$ $p_{3,1}$ $p_{4,1}$ $p_{4,2}$ $p_{5,2}$ $p_{5,3}$ $p_{6,3}$
l^0 l^0 l^1 l^1 l^1 l^2 l^2 l^3 l^3
h^1 h^2 h^2 h^3 h^4 h^4 h^5 h^5 h^6

Score Computation



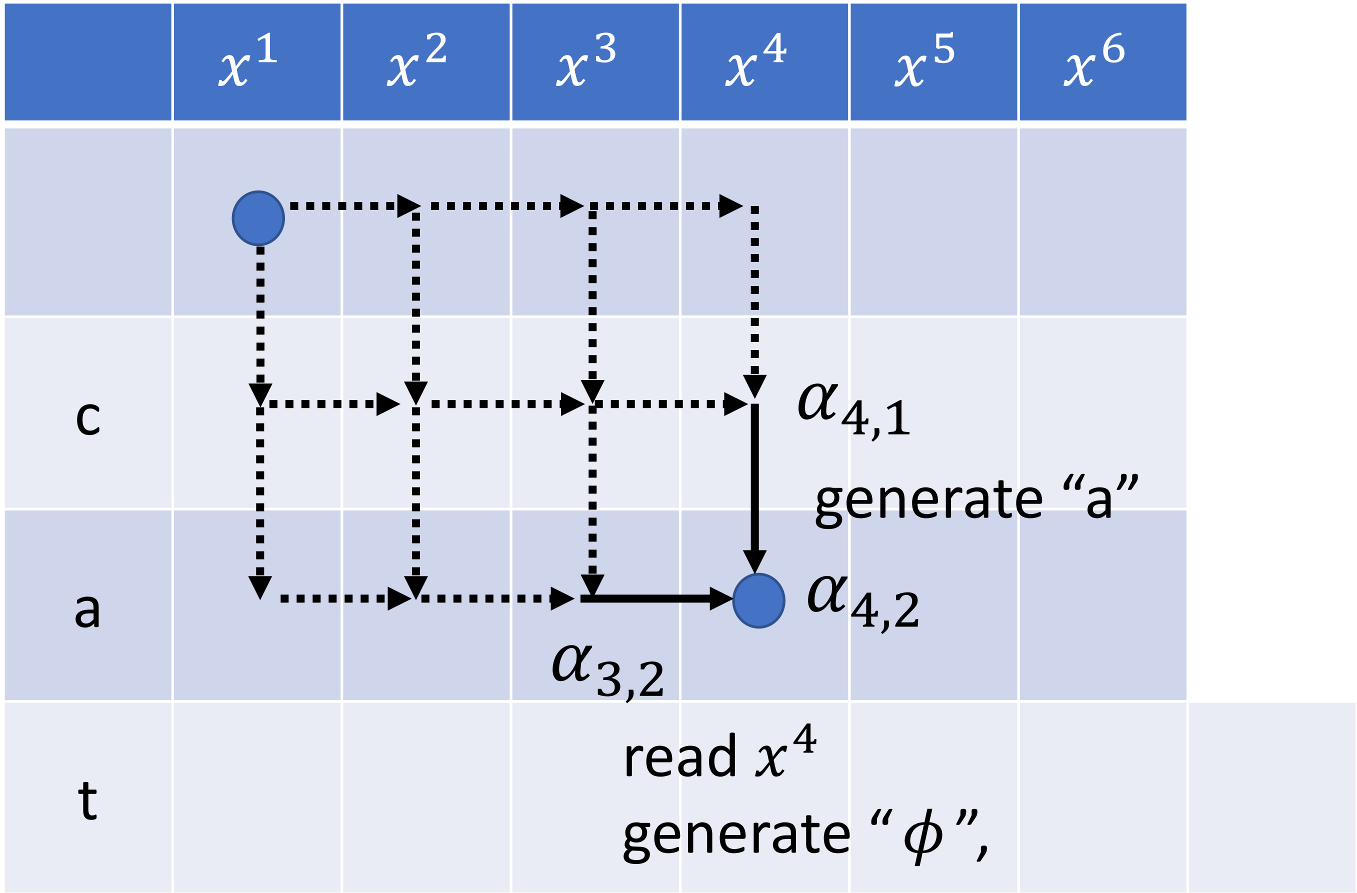
Score Computation



Because ϕ is not considered!

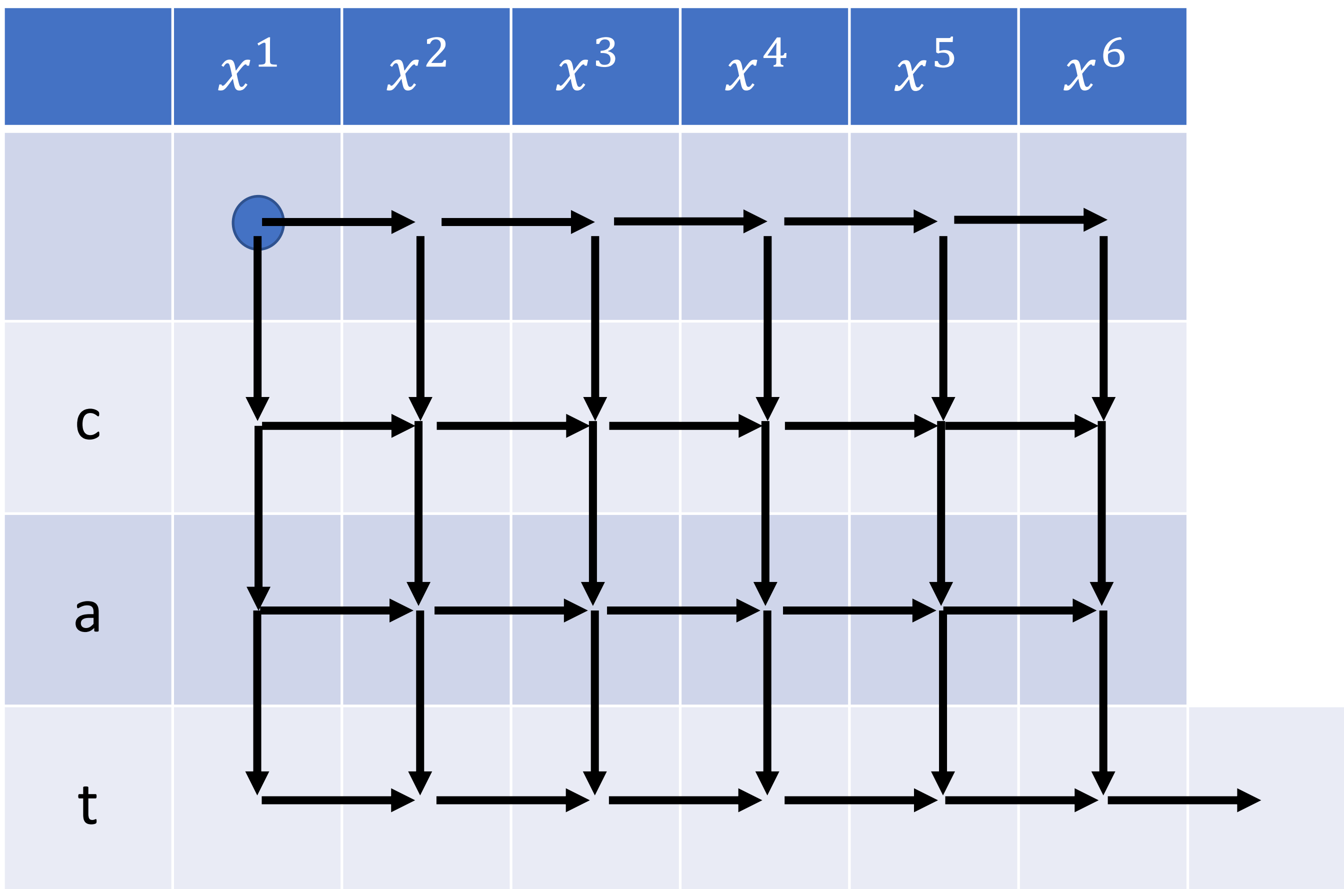
$\alpha_{i,j}$: the summation of the scores of all the alignments that read i-th acoustic features and output j-th tokens

$$\alpha_{4,2} = \alpha_{4,1} p_{4,1}(a) + \alpha_{3,2} p_{3,2}(\phi)$$



$\alpha_{i,j}$: the summation of the scores of all the alignments that read i-th acoustic features and output j-th tokens

$$\alpha_{4,2} = \alpha_{4,1} p_{4,1}(a) + \alpha_{3,2} p_{3,2}(\phi)$$



You can compute summation of the scores of all the alignments.

HMM, CTC, RNN-T

HMM

$$\mathrm{P}_\theta(X|\mathrm{Y}) = \boxed{\sum_{h \in align(Y)} P(X|h)}$$

CTC, RNN-T

$$\mathrm{P}_\theta(\mathrm{Y}|X) = \boxed{\sum_{h \in align(Y)} P(h|X)}$$

1. Enumerate all the possible alignments

2. How to sum over all the alignments

3. Training: $\boxed{\theta^* = arg \max_\theta log \mathrm{P}_\theta\left(\hat{Y}|X\right)}$ $\quad \dfrac{\partial P\left(\hat{Y}|X\right)}{\partial \theta} = ?$

4. Testing (Inference, decoding):

$$\mathrm{Y}^* = arg \max_{\mathrm{Y}} log \mathrm{P}_\theta(\mathrm{Y}|X)$$

Training

$$\theta^* = arg\max_{\theta} logP(\hat{Y}|X)$$

$$\begin{tikzcd} & p_{4,1}(a) \arrow[dr] & \\ \theta \arrow[ur] \arrow[r] \arrow[dr] & p_{3,2}(\phi) \arrow[r] & P(\hat{Y}|X) \\ & \cdots\cdots \arrow[ur] & \end{tikzcd}$$

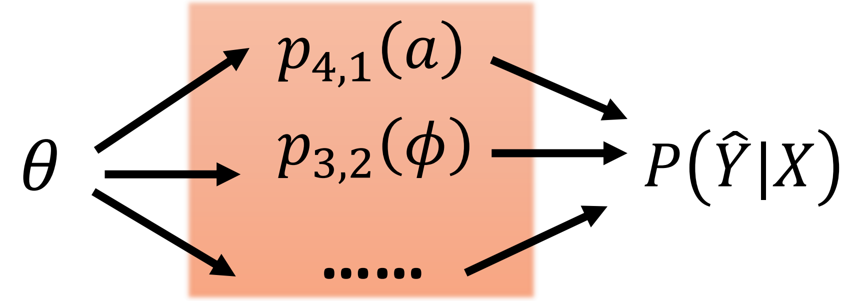
$$P(\hat{Y}|X) = \sum_h P(h|X)$$

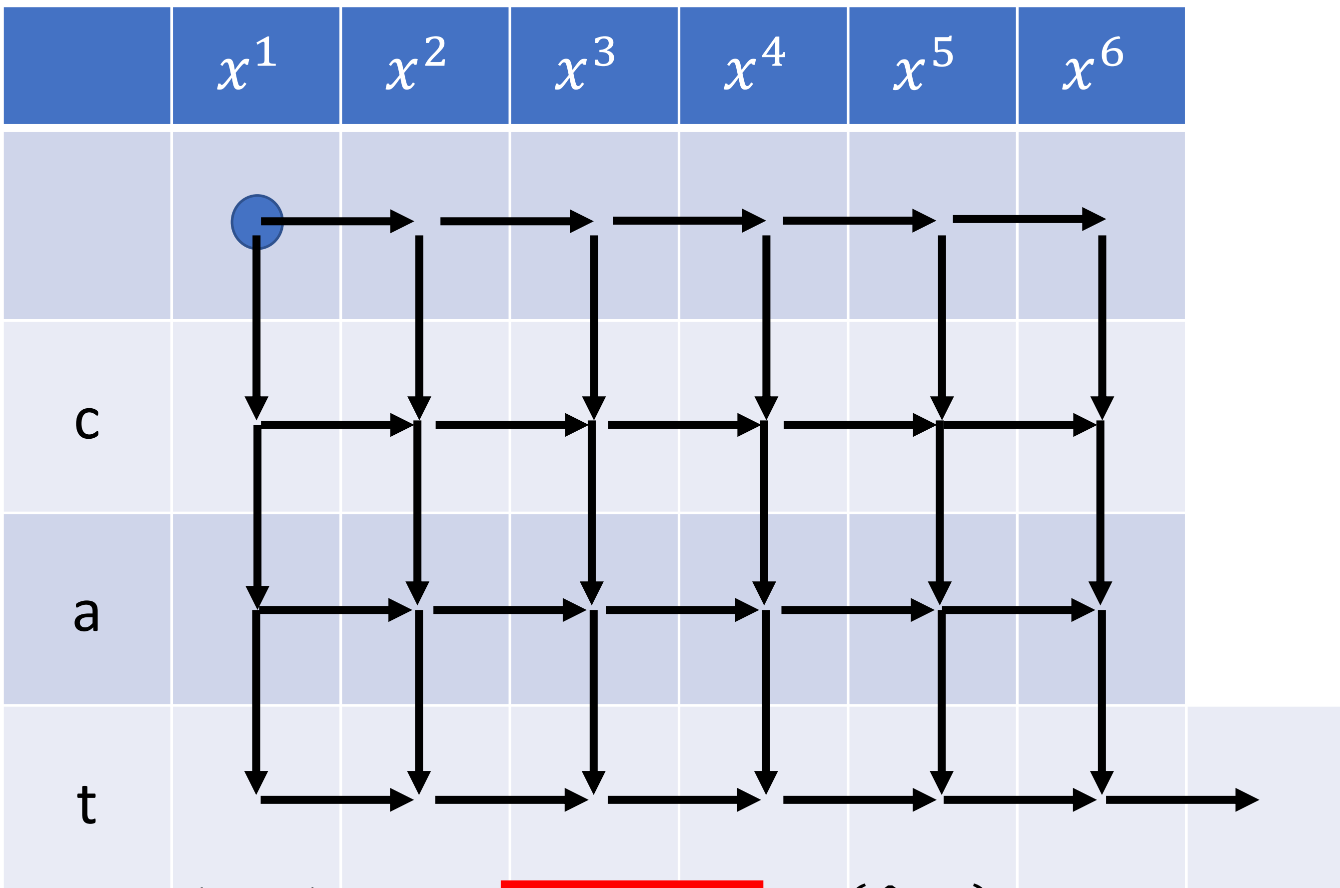
ϕ c ϕ ϕ a ϕ t ϕ ϕ

$$p_{1,0}(\phi) \quad p_{2,0}(c) \quad p_{2,1}(\phi) \quad p_{3,1}(\phi) \quad p_{4,1}(a) \quad p_{4,2}(\phi) \quad p_{5,2}(t) \quad p_{5,3}(\phi) \quad p_{6,3}(\phi)$$

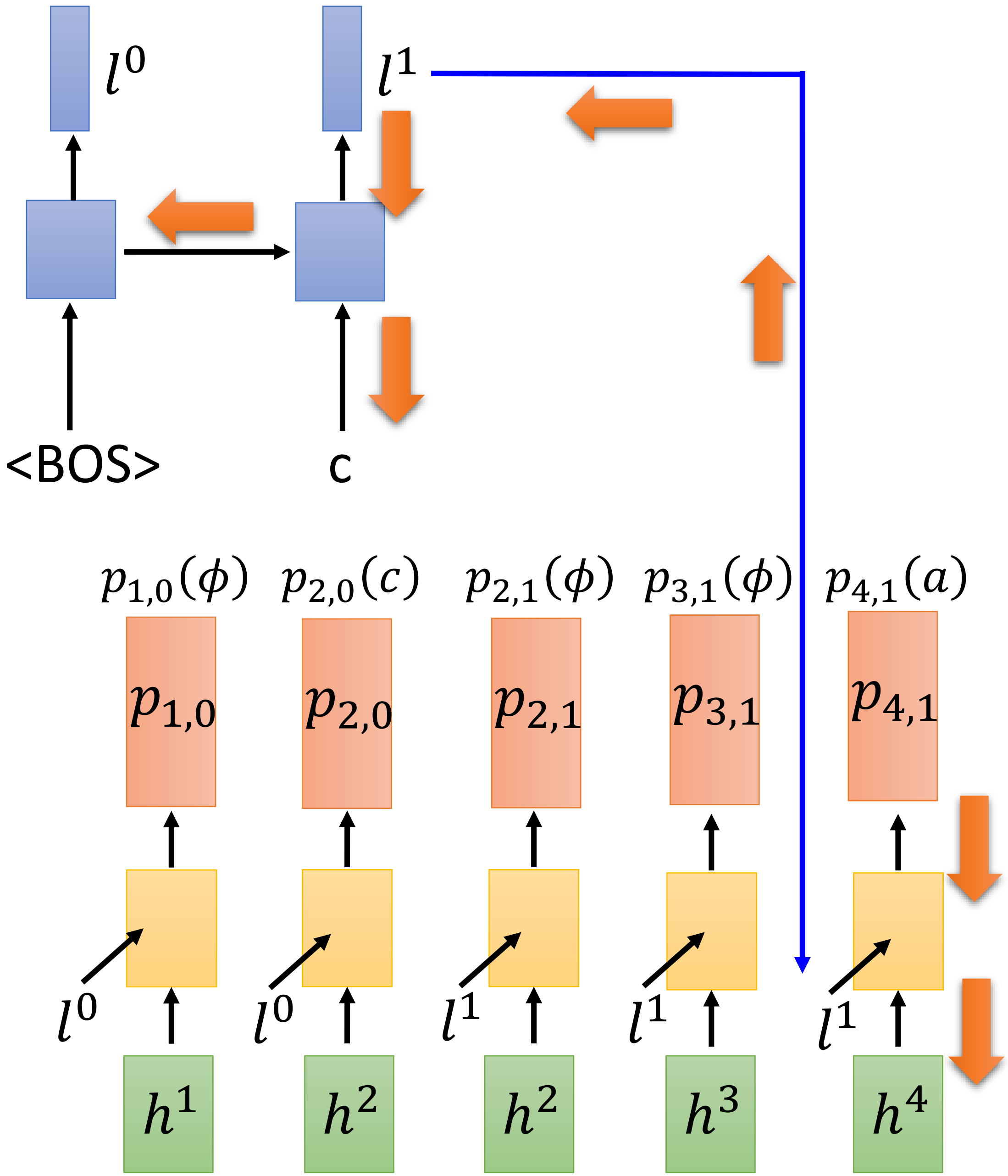
$$\frac{\partial P(\hat{Y}|X)}{\partial \theta} = ? \qquad \frac{\partial p_{4,1}(a)}{\partial \theta}\frac{\partial P(\hat{Y}|X)}{\partial p_{4,1}(a)} + \frac{\partial p_{3,2}(\phi)}{\partial \theta}\frac{\partial P(\hat{Y}|X)}{\partial p_{3,2}(\phi)} + \cdots$$

Each arrow is a component





$$\frac{\partial P(\hat{Y}|X)}{\partial \theta} = ? \quad \boxed{\frac{\partial p_{4,1}(a)}{\partial \theta}} \frac{\partial P(\hat{Y}|X)}{\partial p_{4,1}(a)} + \frac{\partial p_{3,2}(\phi)}{\partial \theta} \frac{\partial P(\hat{Y}|X)}{\partial p_{3,2}(\phi)} + \cdots$$



$$\frac{\partial p_{4,1}(a)}{\partial \theta} = ?$$

Backpropagation (through time)

$$\frac{\partial P(\hat{Y}|X)}{\partial \theta} = ? \quad \frac{\partial p_{4,1}(a)}{\partial \theta} \boxed{\frac{\partial P(\hat{Y}|X)}{\partial p_{4,1}(a)}} + \frac{\partial p_{3,2}(\phi)}{\partial \theta} \frac{\partial P(\hat{Y}|X)}{\partial p_{3,2}(\phi)} + \cdots$$

$$P(\hat{Y}|X) = \sum_{h \text{ with } p_{4,1}(a)} \boxed{P(h|X)} + \sum_{h \text{ without } p_{4,1}(a)} P(h|X)$$

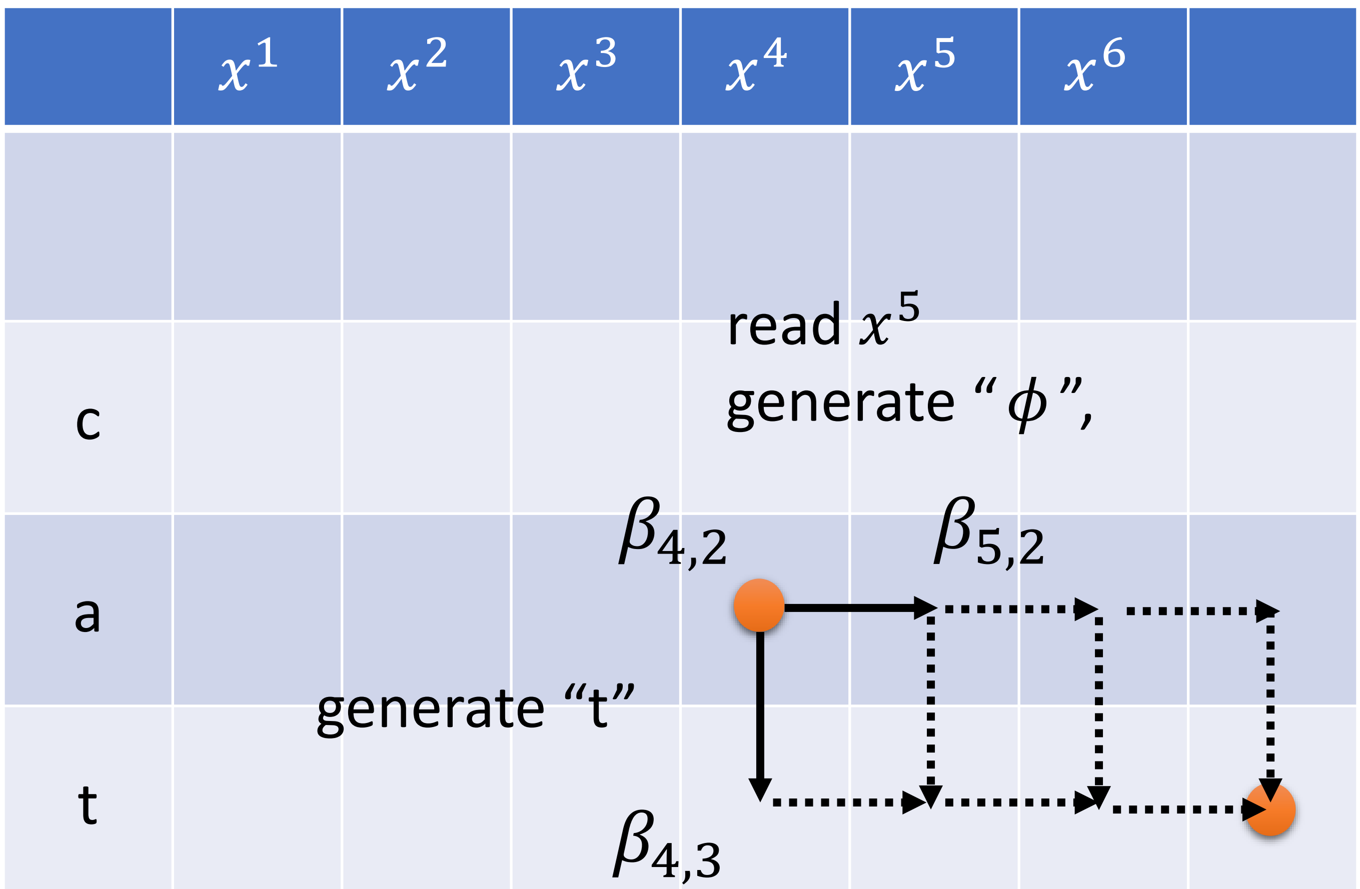
$$p_{4,1}(a) \times other$$

$$\frac{\partial P(\hat{Y}|X)}{\partial p_{4,1}(a)} = \sum_{h \text{ with } p_{4,1}(a)} other = \sum_{h \text{ with } p_{4,1}(a)} \frac{P(h|X)}{p_{4,1}(a)}$$

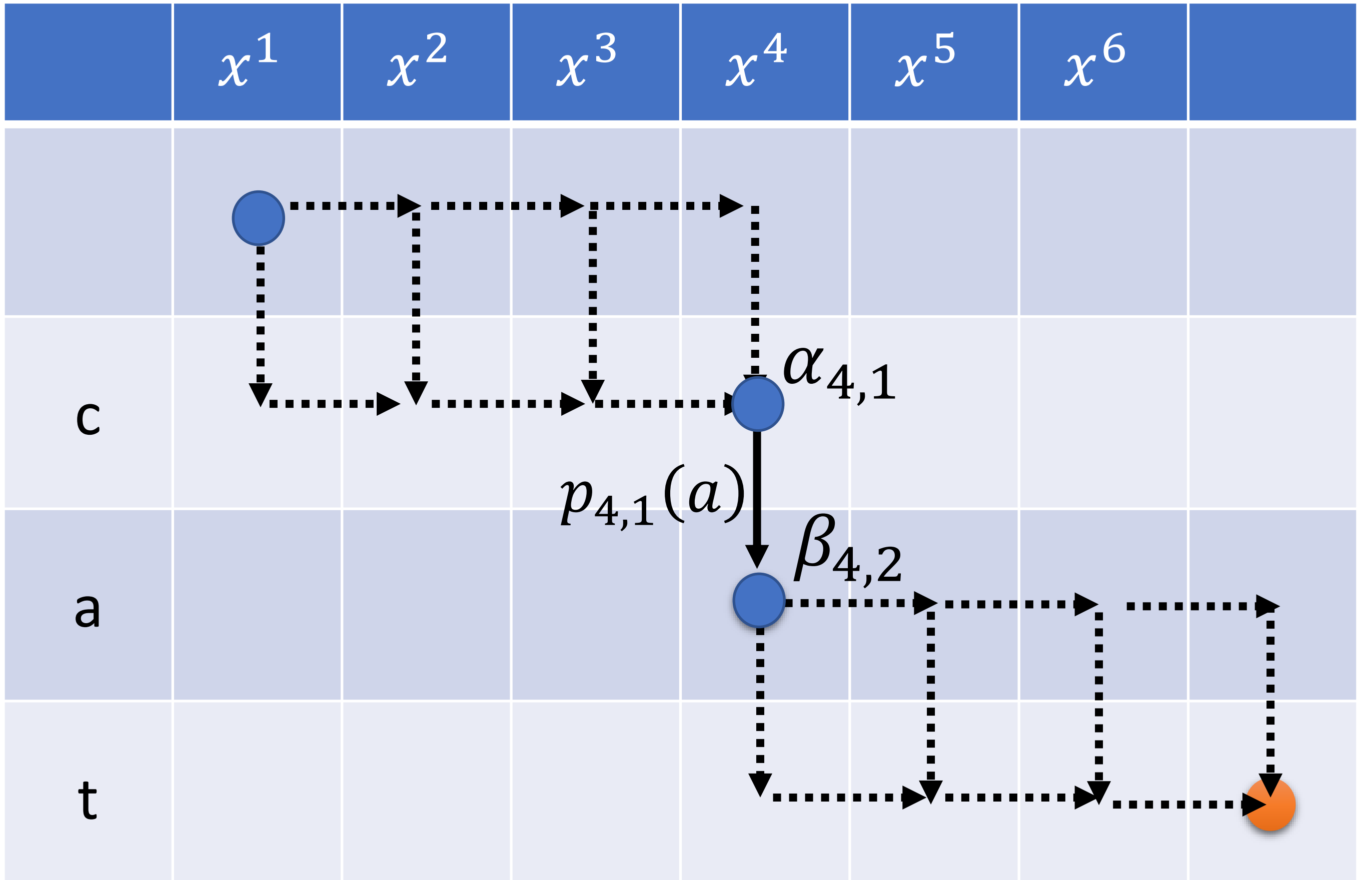
$$= \frac{1}{p_{4,1}(a)} \sum_{h \text{ with } p_{4,1}(a)} P(h|X)$$

$\beta_{i,j}$: the summation of the score of all the alignments staring from i-th acoustic features and j-th tokens

$$\beta_{4,2} = \beta_{4,3}p_{4,2}(t) + \beta_{5,2}p_{4,2}(\phi)$$



$$\frac{\partial P(\hat{Y}|X)}{\partial p_{4,1}(a)} = \frac{1}{p_{4,1}(a)} \boxed{\sum_{a \text{ with } p_{4,1}(a)} P(a|X)} \; \alpha_{4,1}\, p_{4,1}(a) \beta_{4,2}$$



HMM, CTC, RNN-T

HMM

$$\mathrm{P}_\theta(X|\mathrm{Y}) = \boxed{\sum_{\underline{h \in align(Y)}} P(X|h)}$$

CTC, RNN-T

$$\mathrm{P}_\theta(\mathrm{Y}|X) = \boxed{\sum_{\underline{h \in align(Y)}} P(h|X)}$$

1. Enumerate all the possible alignments

2. How to sum over all the alignments

3. Training: $\boxed{\theta^* = arg \max_\theta log \mathrm{P}_\theta\left(\hat{Y}|X\right)}$ $\quad \dfrac{\partial P\left(\hat{Y}|X\right)}{\partial \theta} = ?$

4. Testing (Inference, decoding):

$$\mathrm{Y}^* = arg \max_\mathrm{Y} log \mathrm{P}_\theta(\mathrm{Y}|X)$$

Summary

	LAS	CTC	RNN-T
Decoder	Not independent	independent	Not independent
Alignment	Not explicit (Soft alignment)	Yes	Yes
Training	Just train it	Sum over alignment	Sum over alignment
Streaming	No	Yes	Yes

Reference

- [Yu, et al., INTERSPEECH’16] Dong Yu, Wayne Xiong, Jasha Droppo, Andreas Stolcke , Guoli Ye, Jinyu Li , Geoffrey Zweig, Deep Convolutional Neural Networks with Layer-wise Context Expansion and Attention, INTERSPEECH, 2016
- [Saon, et al., INTERSPEECH’17] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, Phil Hall, English Conversational Telephone Speech Recognition by Humans and Machines, INTERSPEECH, 2017