NATURAL LANGUAGE PROCESSING

WITH DEEP LEARNING

University of Guilan

Lecturer: Javad PourMostafa

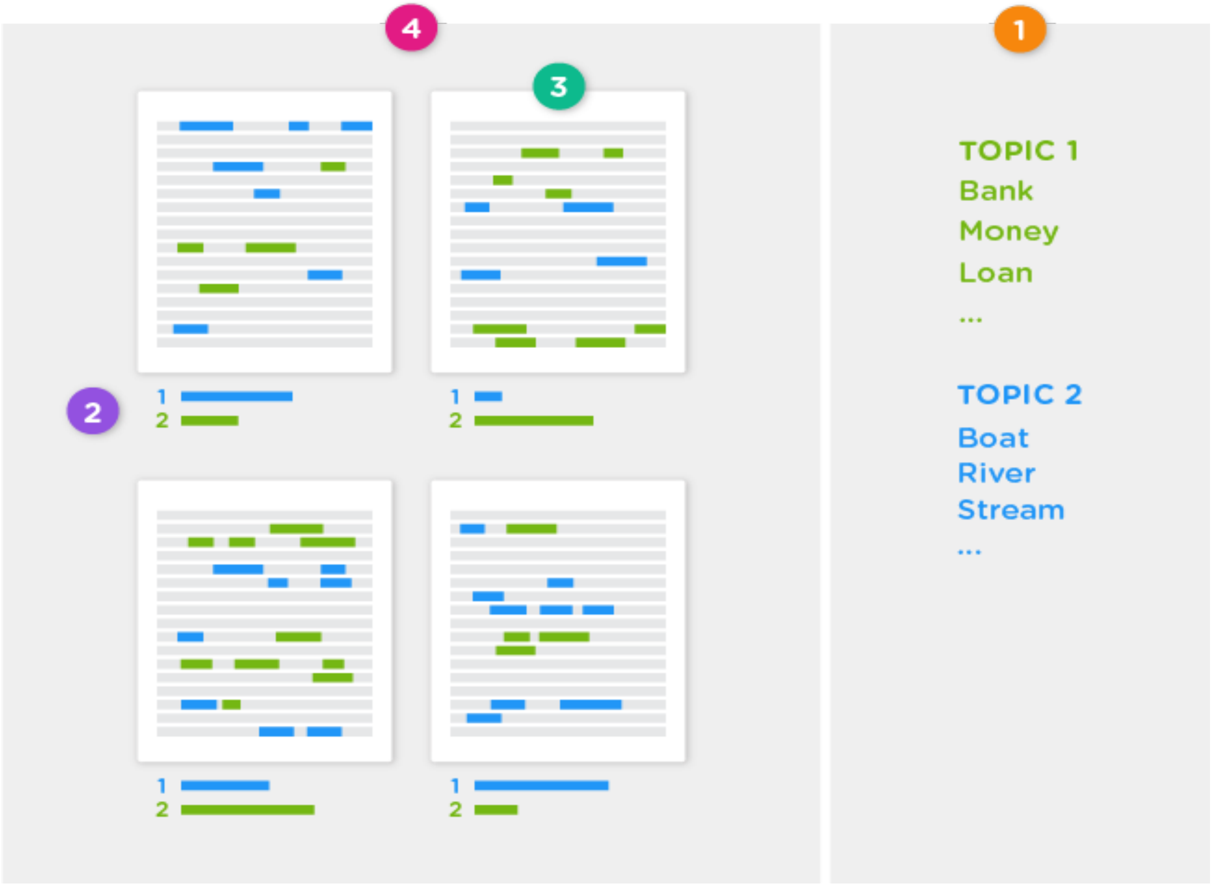
NLP981

Topic Modeling

TOPIC MODELING

- Topic Modeling is method for identifying 1 topics and their 2 distributions across the 3 documents in a 4 corpus.

- *Where do we need that?*
  - Weblogs categories
  - *Chat-bots (text generation based on topics)*
  - *Recommender Systems*
  - …



TOPIC MODELING

- Given:
  - Collection of texts as bags-of-words:
  - n_{wd} is a count of the words w in the document d

- Find:
  - Probabilities of word in topics:
  - $\emptyset_{wt} = p(w|t)$
  - Probabilities of topics in documents:
  - $\theta_{td} = p(t|d)$

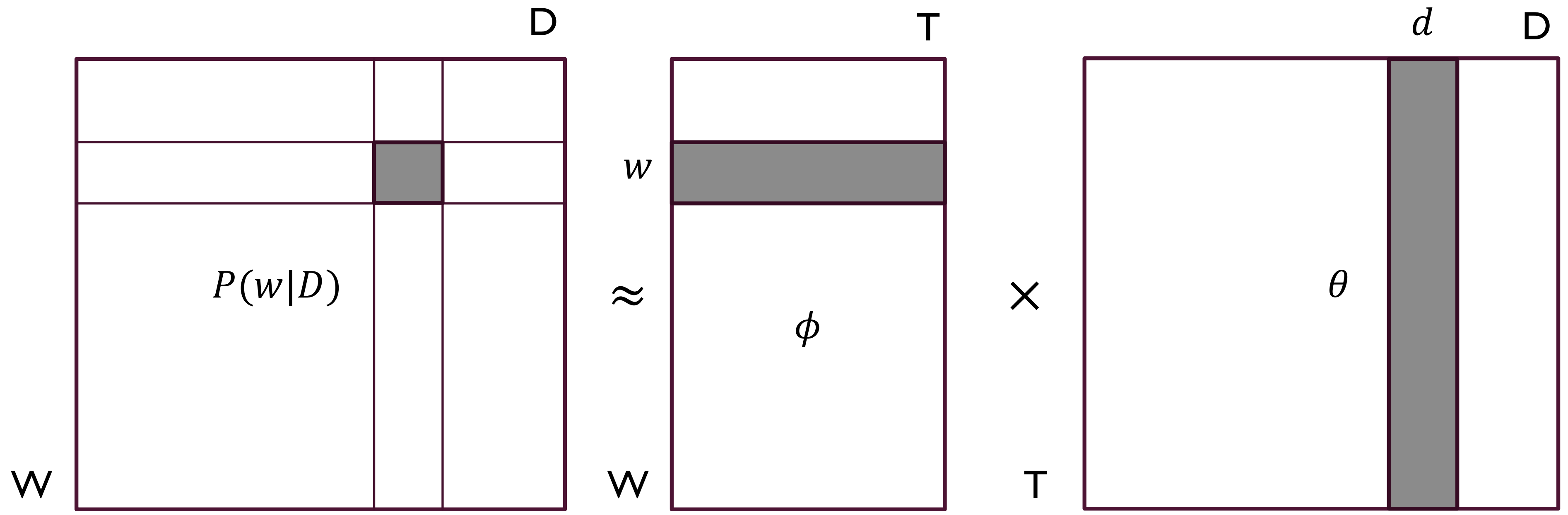
Topic Models:

PLSA, LDA and …

TOPIC MODEL: PLSA

- ***Probabilistic Latent Semantic Analysis*** (Thomas Hofmann)

$$\boldsymbol{P(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \emptyset_{wt}\, \theta_{td}}$$



AN EXAMPLE OF TOPIC MODELING WITH PLSA

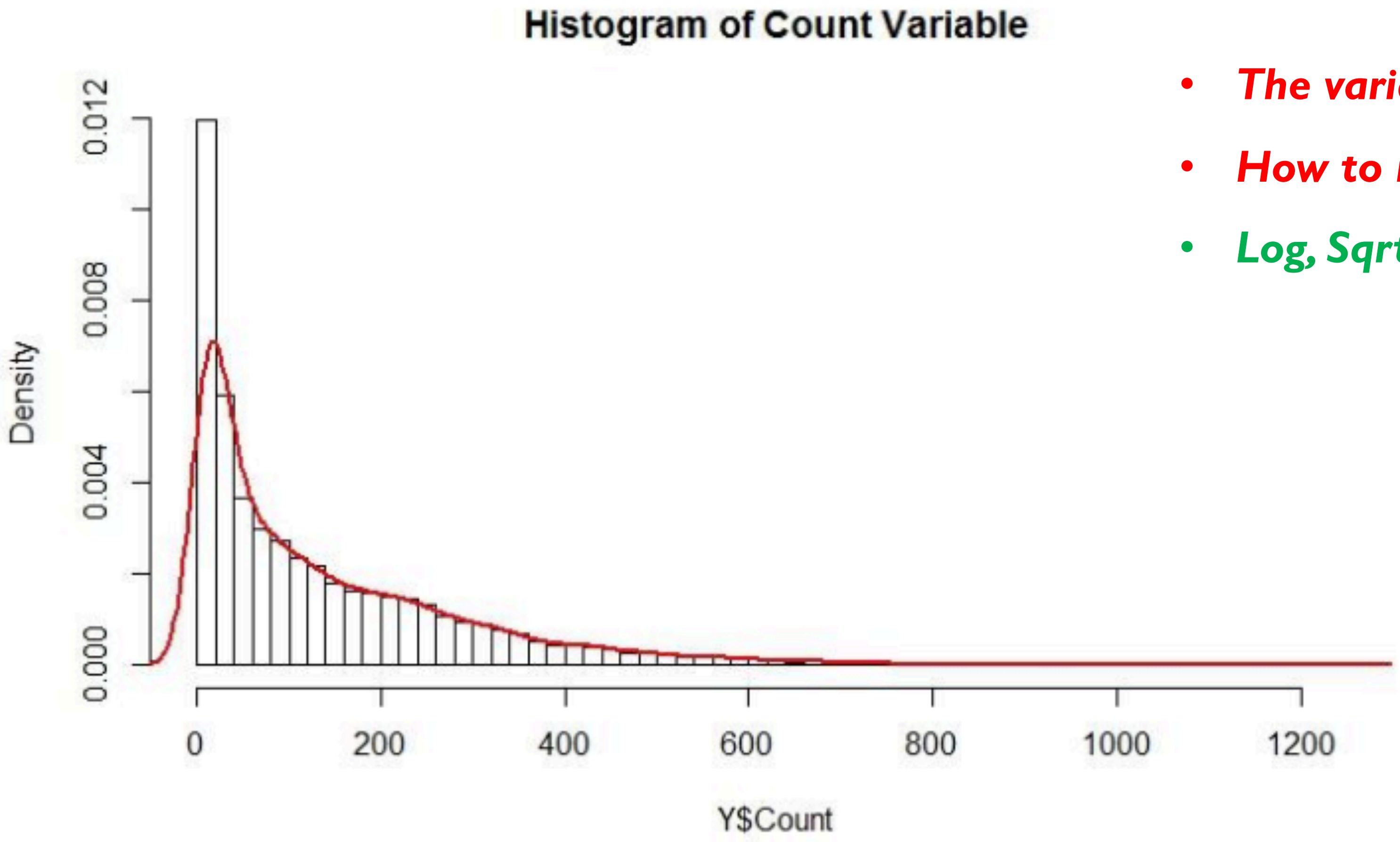
- Pooh rubbed his nose again, and said that he hadn't thought of that. And then he brightened up, and said that, if it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he wouldn't see the Very Deep Pit until he was half-way down....

- $P(w = raining|\boldsymbol{t}) = \frac{n_{wt}}{\sum_w n_{wt}} = \frac{1}{4}$
- $P(t = \boldsymbol{r}|d) = \frac{n_{td}}{\sum_t n_{td}} = \frac{4}{54}$

- But we have just plain texts in real worlds
- Topics are hidden

How to deal with the issue?

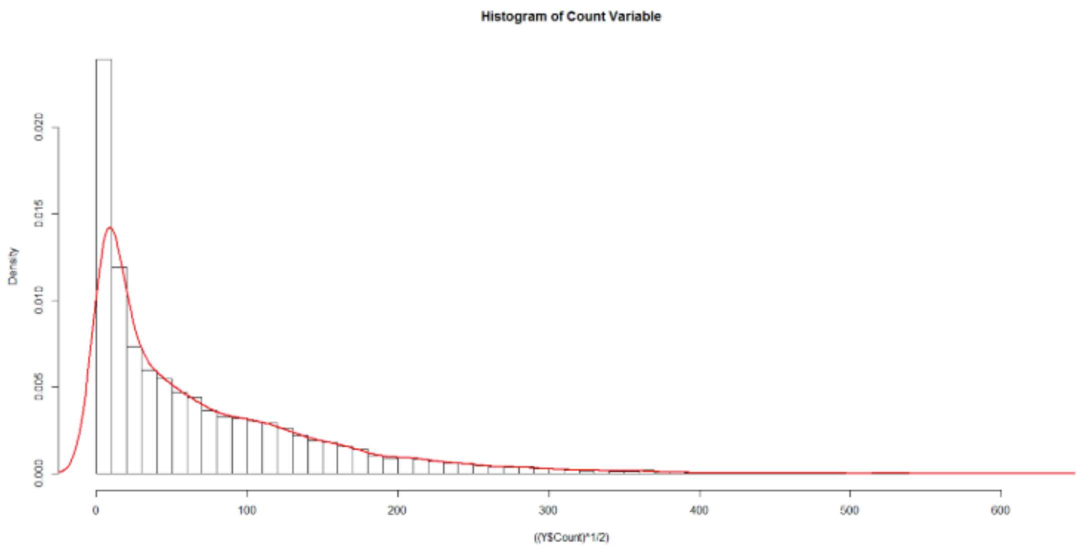
PROBLEM OF PROBABILITY DENSITY ESTIMATION (1)



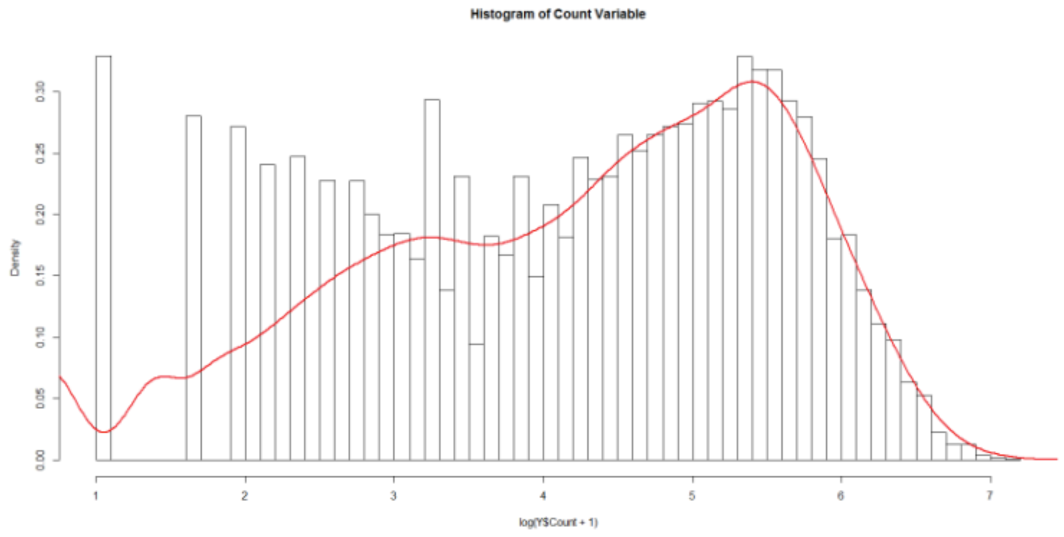
- *The variable is not normally distributed*
- *How to model such a variable?*
- *Log, Sqrt, …*

PROBLEM OF PROBABILITY DENSITY ESTIMATION (2)

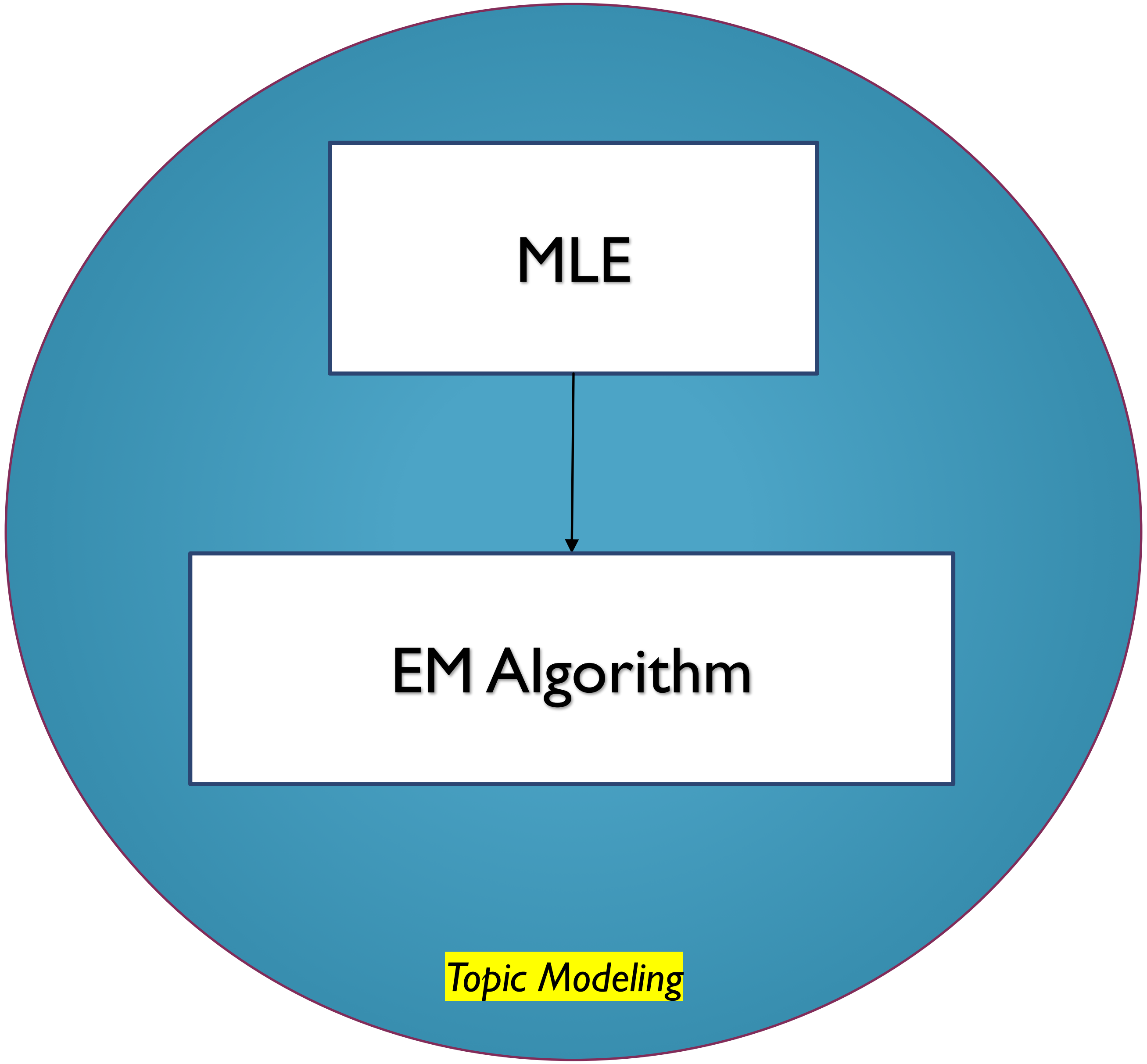
With Square Root Transformation:



With Log transformation:



- This is where **Maximum Likelihood Estimation (MLE)** has such a major advantage


MLE
EM Algorithm
Topic Modeling

MAXIMUM LIKELIHOOD ESTIMATION

- **Given Some Data:** $(x_1, x_2, x_3, x_4, \dots, x_n)$
- **MLE defines a parameter called theta** θ
- **we wish to maximize the probability of observing data from join probability**
  - $P(X \mid Theta) = P(\, x_1, x_2, \dots, x_n \mid Theta)$
- **This conditional probability is referred to Likelihood**
  - $L(X \mid Theta)$
- ***We'd like to maximize it. Thus,***
  - $Maximize\ L(X \mid Theta)$

MAXIMUM LIKELIHOOD ESTIMATION (CONT'D)

- ***Don't forget it's a join probability distribution:***
  - $\boldsymbol{Product\ i\ to\ n\ P(x_i \mid Theta)}$
- *Note: Multiplying many small probabilities together can be numerically unstable in practice*
  - $\boldsymbol{Sum\ i\ to\ n\ P(x_i \mid Theta) = Sum\ i\ to\ n\ log(P(x_i \mid Theta))}$
- How to map it into a **topic model or Machine Learning algorithms**?

$$\boldsymbol{Maxmimize\ sum\ i\ to\ n\ \log(P(\ X \mid h))}$$

USING MLE IN PLSA

A. $P(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \emptyset_{wt}\, \theta_{td}$

B. $Maxmimize\ sum\ i\ to\ n \log(P(\,X \mid h))$

- Parameters: $\emptyset_{wt}, \theta_{td}$ *that we like to maximize*
- *Our parameters are non negative cause they are probabilities, thus,*
  - $\emptyset_{wt} \geq 0\ and\ \theta_{td} \geq 0$
  - $\sum_{w \in W} \emptyset_{wt} = 1\ \ and\ \sum_{t \in T} \theta_{td} = 1$ ***(normalization)***
- Computing of $\mathrm{Log} \sum_{t \in T} \emptyset_{wt}\, \theta_{td}$ is impossible cause of latent variables!

USING EXPECTATION-MAXIMIZATION IN PLSA

- ***EM algorithm:***
  - ***an approach for performing MLE in the presence of latent variables***
- ***Two coupled steps:***

i. ***E-Step: Estimate the missing variables in the dataset***

ii. ***M-Step: Maximize the parameters of the model in the presence of the data***

E-STEP AND M-STEP IN PLSA

E-Step:

$$P(t \mid d, w) = \frac{P(w|t)p(t|d)}{p(w|d)} = \frac{\emptyset_{wt}\theta_{td}}{\sum_{s \in T} \emptyset_{ws}\theta_{sd}}$$

M-Step (Update for parameters):

$$\emptyset_{wt} = \frac{n_{wt}}{\sum_w n_{wt}} \qquad n_{wt} = \sum_d n_{dw}\, p(t|d,w)$$

$$\theta_{td} = \frac{n_{td}}{\sum_t n_{td}} \qquad n_{td} = \sum_w n_{dw}\, p(t|d,w)$$

Do the steps until convergence!

AN EXAMPLE OF EM-ALGORITHM UPDATE

- ***Plain Text:*** *If it were raining already, the Heffalump would be looking at the sky wondering if it would clear up, and so he would not see the Very Deep Pit until he was half-way down…*
- ***1st assumption: 3 topics in our model***
- ***2nd assumption: random initialization for parameters*** $(\theta, \emptyset)$

$\emptyset$ ***matrix***

	Topic 1	Topic 2	Topic 3
raining	0.01	0.1	0.05
would	0.1	0.2	0.1
…	…	…	…

$\boldsymbol{\theta}$ ***matrix***

	Document
Topic 1	0.1
Topic 2	0.5
Topic 3	0.4

AN EXAMPLE OF EM-ALGORITHM UPDATE

$\emptyset$ ***matrix***

	Topic 1	Topic 2	Topic 3
raining	0.01	0.1	0.05
would	0.1	0.2	0.1
…	…	…	…

θ ***matrix***

	Document
Topic 1	0.1
Topic 2	0.5
Topic 3	0.4

$$E-step\text{: } P(t \mid d, w) = \frac{P(w|t)p(t|d)}{p(w|d)} = \frac{\emptyset_{wt}\theta_{td}}{\sum_{s\in T}\emptyset_{ws}\theta_{sd}} = \frac{0.2\times 0.5}{(0.1\times 0.1)+(0.2\times 0.5)+(0.1\times 0.4)} = \frac{0.1}{0.15} = 0.66$$

$$M-Step\text{: } \emptyset_{wt} = \frac{n_{wt}}{\sum_w n_{wt}} \qquad n_{wt} = \sum_d n_{dw}\, p(t|d,w) \quad = 3\times 0.66 \cong 2.0$$

MORE INFORMATION

- ***How many parameters does PLSA topic model have?***
  - ***Vocabulary size |w|, Documents number |D|, Length of Corpus |N| and number of topics |T|***
- $|\boldsymbol{W}| \,.\, |\boldsymbol{T}| \;+\; |\boldsymbol{T}| \,.\, |\boldsymbol{D}|$