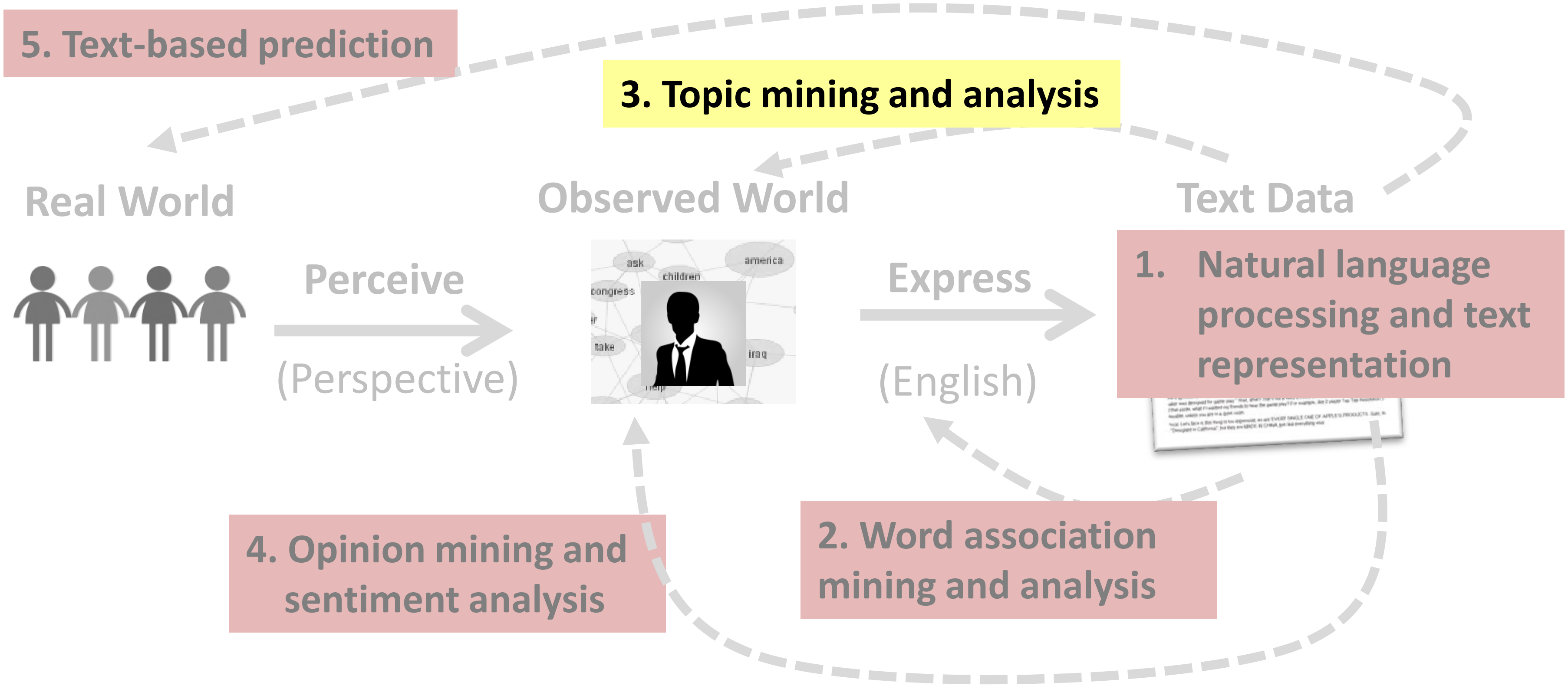
Probabilistic Topic Models: Expectation-Maximization Algorithm

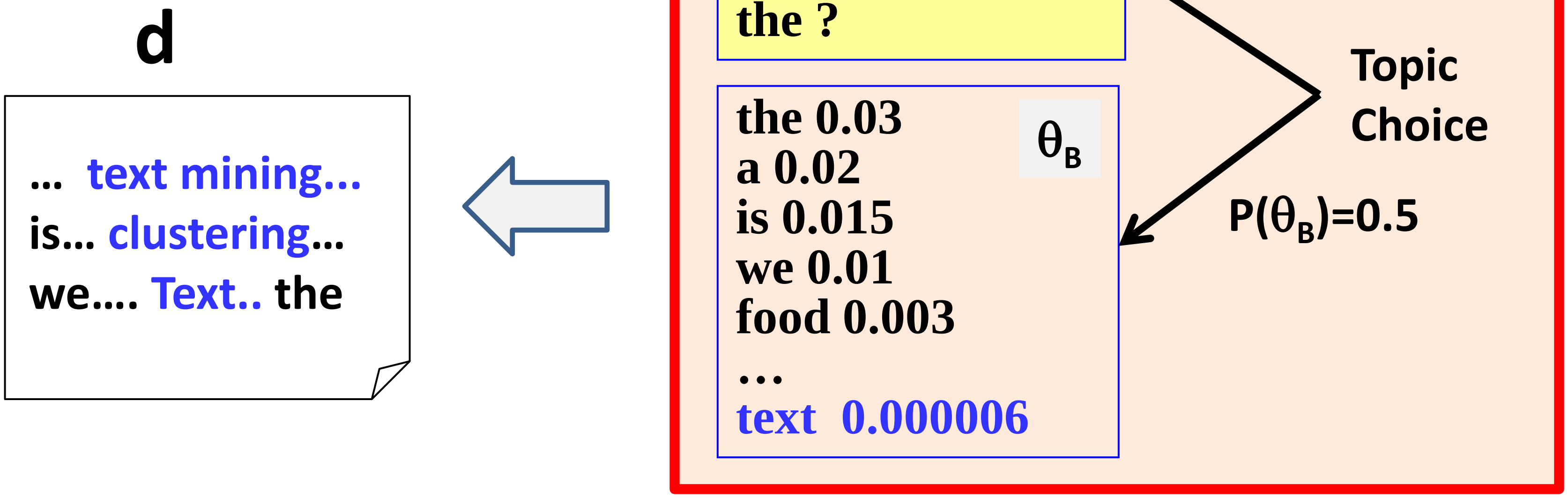
ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Probabilistic Topic Models: Expectation-Maximization (EM) Algorithm



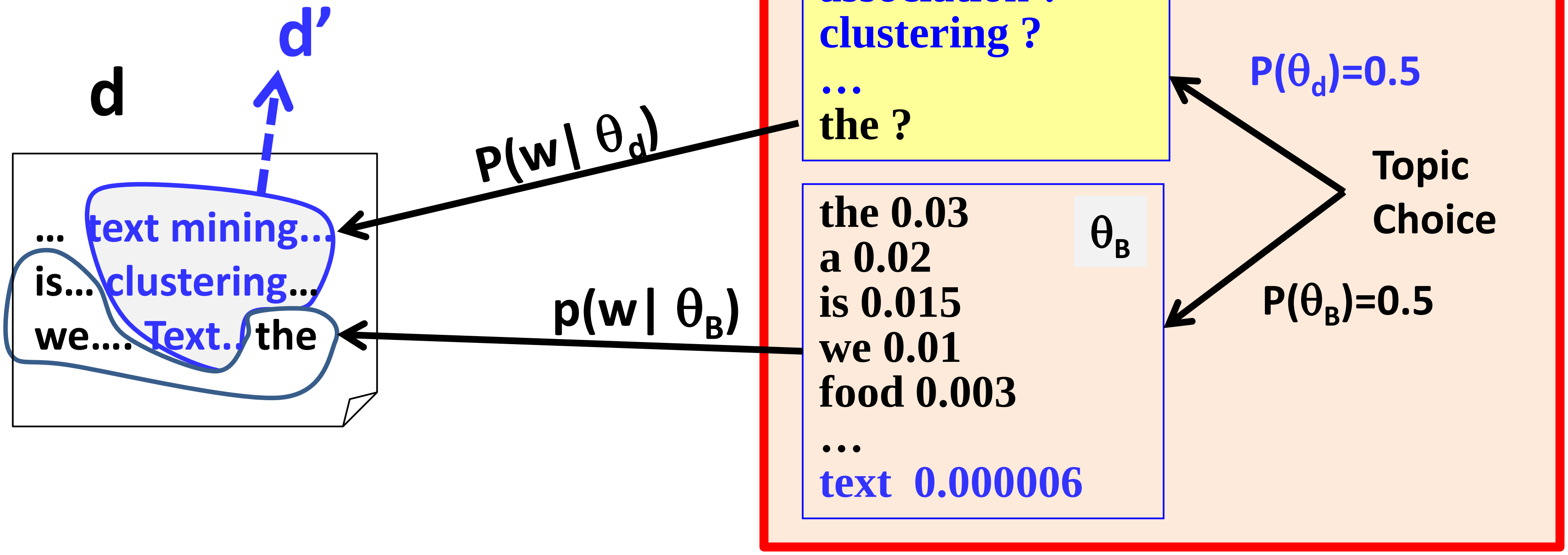
Estimation of One Topic: $\mathbf{P(w|\ \theta_d)}$

How to set θ_d to maximize $p(d|\Lambda)$?
(all other parameters are known)



If we know which word is from which distribution…

$$p(w_i \mid \theta_d) = \frac{c(w_i, d')}{\sum_{w' \in V} c(w', d')}$$



Given all the parameters, infer the distribution a word is from…

Is "text" more likely from θ_d or θ_B ?

From θ_d (Z=0)?

$p(\theta_d)p(\text{"text"}|\theta_d)$

From θ_B (Z=1)?

$p(\theta_B)p(\text{"text"}|\theta_B)$

$P(w|\theta_d)$

$p(w|\theta_B)$

θ_d

text 0.04
mining 0.035
association 0.03
clustering 0.005
…
the 0.000001

θ_B

the 0.03
a 0.02
is 0.015
we 0.01
food 0.003
…
text 0.000006

$p(\theta_d)+p(\theta_B)=1$

$P(\theta_d)=0.5$

Topic Choice

$P(\theta_B)=0.5$

$$p(z=0|w=\text{"text"}) = \frac{p(\theta_d)p(\text{"text"}|\theta_d)}{p(\theta_d)p(\text{"text"}|\theta_d)+p(\theta_B)p(\text{"text"}|\theta_B)}$$

The Expectation-Maximization (EM) Algorithm

Hidden Variable:
$z \in \{0, 1\}$

	z
the	**1**
paper	**1**
presents	**1**
a	**1**
text	**0**
mining	**0**
algorithm	**0**
for	**1**
clustering	**0**
...	**...**

Initialize $p(w|\theta_d)$ with random values.
Then iteratively improve it using E-step & M-step.
Stop when likelihood doesn't change.

$$p^{(n)}(z=0\,|\,w) = \frac{p(\theta_d)p^{(n)}(w\,|\,\theta_d)}{p(\theta_d)p^{(n)}(w\,|\,\theta_d) + p(\theta_B)p(w\,|\,\theta_B)}$$

E-step

How likely w is from θ_d

$$p^{(n+1)}(w\,|\,\theta_d) = \frac{c(w,d)p^{(n)}(z=0\,|\,w)}{\sum_{w'\in V} c(w',d)p^{(n)}(z=0\,|\,w')}$$

M-step

EM Computation in Action

E-step $$p^{(n)}(z=0\mid w)=\frac{p(\theta_d)p^{(n)}(w\mid\theta_d)}{p(\theta_d)p^{(n)}(w\mid\theta_d)+p(\theta_B)p(w\mid\theta_B)}$$

M-step $$p^{(n+1)}(w\mid\theta_d)=\frac{c(w,d)p^{(n)}(z=0\mid w)}{\sum_{w'\in V}c(w',d)p^{(n)}(z=0\mid w')}$$

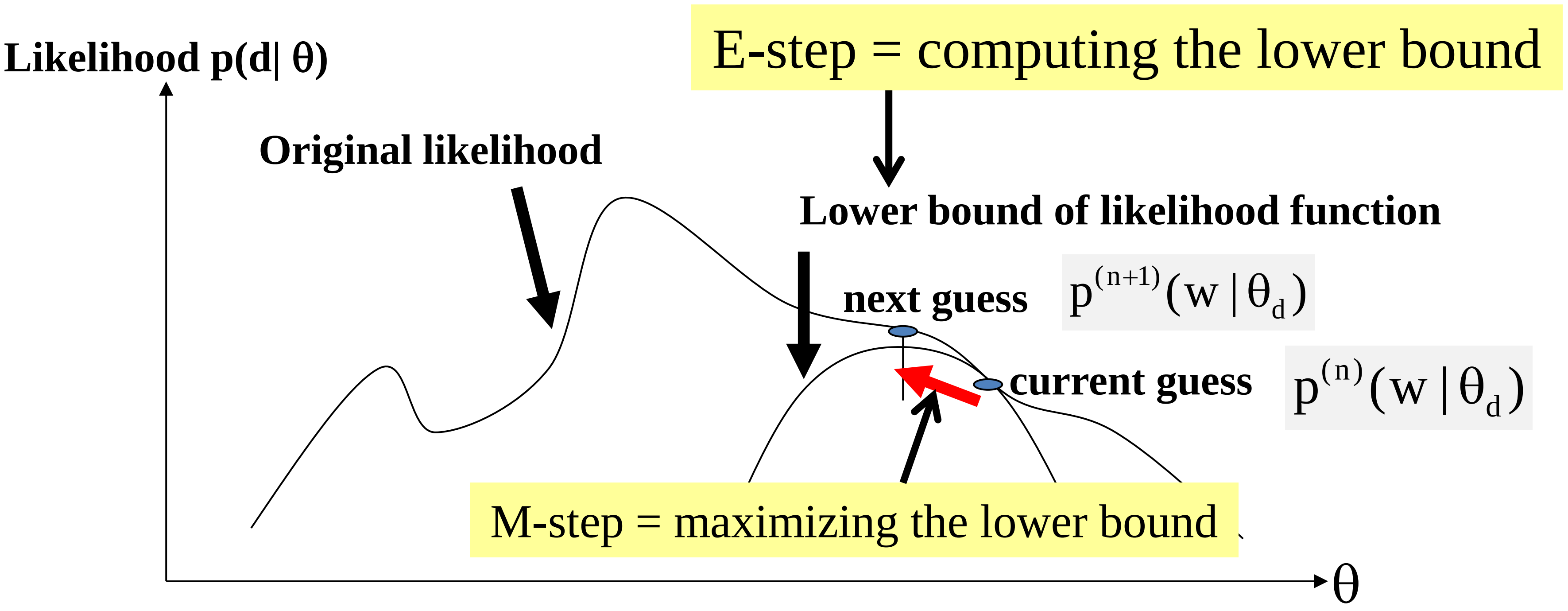
Assume
$p(\theta_d)=p(\theta_B)=0.5$
and $p(w|\theta_B)$ is known

Word	#	$p(w\|\theta_B)$	Iteration 1		Iteration 2		Iteration 3	
			$P(w\|\theta)$	$p(z=0\|w)$	$P(w\|\theta)$	$P(z=0\|w)$	$P(w\|\theta)$	$P(z=0\|w)$
The	4	0.5	**0.25**	0.33	**0.20**	0.29	**0.18**	0.26
Paper	2	0.3	**0.25**	0.45	**0.14**	0.32	**0.10**	0.25
Text	4	0.1	**0.25**	0.71	**0.44**	0.81	**0.50**	0.93
Mining	2	0.1	**0.25**	0.71	**0.22**	0.69	**0.22**	0.69
Log-Likelihood			-16.96		-16.13		-16.02	

Likelihood increasing

“By products”: Are they also useful?

EM As Hill-Climbing ➔ Converge to Local Maximum



Summary

- Expectation-Maximization (EM) algorithm
  - General algorithm for computing ML estimate of mixture models
  - Hill-climbing, so can only converge to a local maximum (depending on initial points)
- E-step: “augment” data by predicting values of useful hidden variables
- M-step: exploit the “augmented data” to improve estimate of parameters (“improve” is guaranteed in terms of likelihood)
- “Data augmentation” is probabilistic ➔ Split counts of events probabilistically