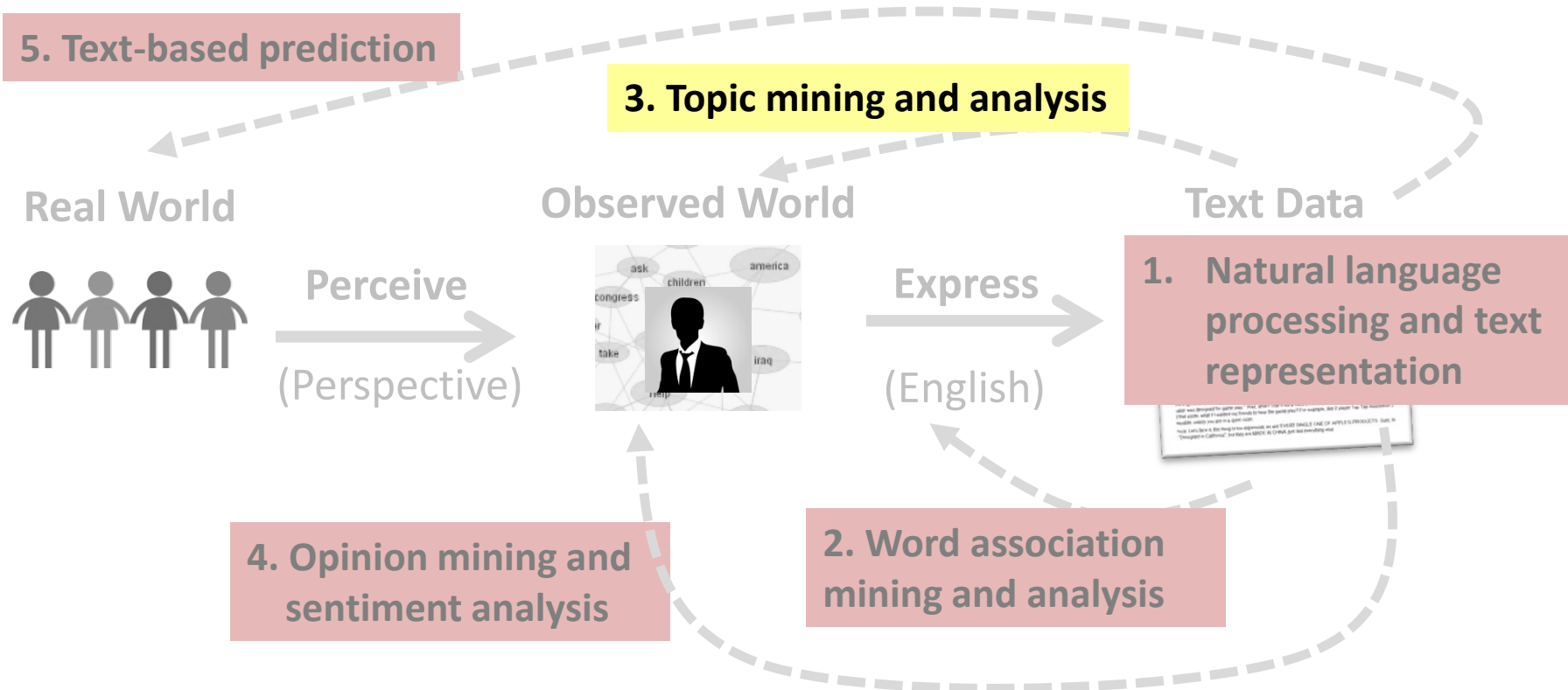# Latent Dirichlet Allocation (LDA)

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

Latent Dirichlet Allocation (LDA)

5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

Perceive

(Perspective)

Express

(English)

ask   america

children

congress

take   iraq

help

1. **Natural language processing and text representation**

4. Opinion mining and sentiment analysis

2. Word association mining and analysis

Extensions of PLSA

- PLSA with prior knowledge ➔ User-controlled PLSA
- PLSA as a generative model ➔ Latent Dirichlet Allocation

PLSA with Prior Knowledge

- Users may have expectations about which topics to analyze:
  - We expect to see "retrieval models" as a topic in IR
  - We want to see aspects such as "battery" and "memory" for opinions about a laptop
- Users may have knowledge about what topics are (or are NOT) covered in a document
  - Tags = topics ➔ A doc can only be generated using topics corresponding to the tags assigned to the document
- We can incorporate such knowledge as priors of PLSA model

Maximum a Posteriori (MAP) Estimate

$$\Lambda^* = \arg\max_{\Lambda} p(\Lambda)\, p(Data \mid \Lambda)$$

- We may use p(Λ) to encode all kinds of preferences and constraints, e.g.,
  - p(Λ)>0 if and only if one topic is precisely "background": p(w|θ_B)
  - p(Λ)>0 if and only if for a particular doc d, $\pi_{d,3}$=0 and $\pi_{d,1}$=1/2
  - p(Λ) favors a Λ with topics that assign high probabilities to some particular words
- The MAP estimate (with conjugate prior) can be computed using a similar EM algorithm to the ML estimate with smoothing to reflect prior preferences

5

EM Algorithm with Conjugate Prior on p(w| θ_i)

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w \mid \theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)}(w \mid \theta_{j'})}$$

Prior: p(w|θ′$_j$)

battery 0.5
life 0.5

$$p(z_{d,w} = B) = \frac{\lambda_B p(w \mid \theta_B)}{\lambda_B p(w \mid \theta_B) + (1 - \lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)}(w \mid \theta_j)}$$

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j)}{\sum_{j'} \sum_{w \in V} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j')}$$

Pseudo counts of w
from prior θ'

$$p^{(n+1)}(w \mid \theta_j) = \frac{\sum_{d \in C} c(w,d)(1 - p(z_{d,w} = B)) p(z_{d,w} = j) \; \textbf{+µp(w|θ'}_j\textbf{)}}{\sum_{w' \in V} \sum_{d \in C} c(w',d)(1 - p(z_{d,w'} = B)) p(z_{d,w'} = j) \; \textbf{+µ}}$$

What if µ=0? What if µ=+∞?

Sum of all pseudo counts

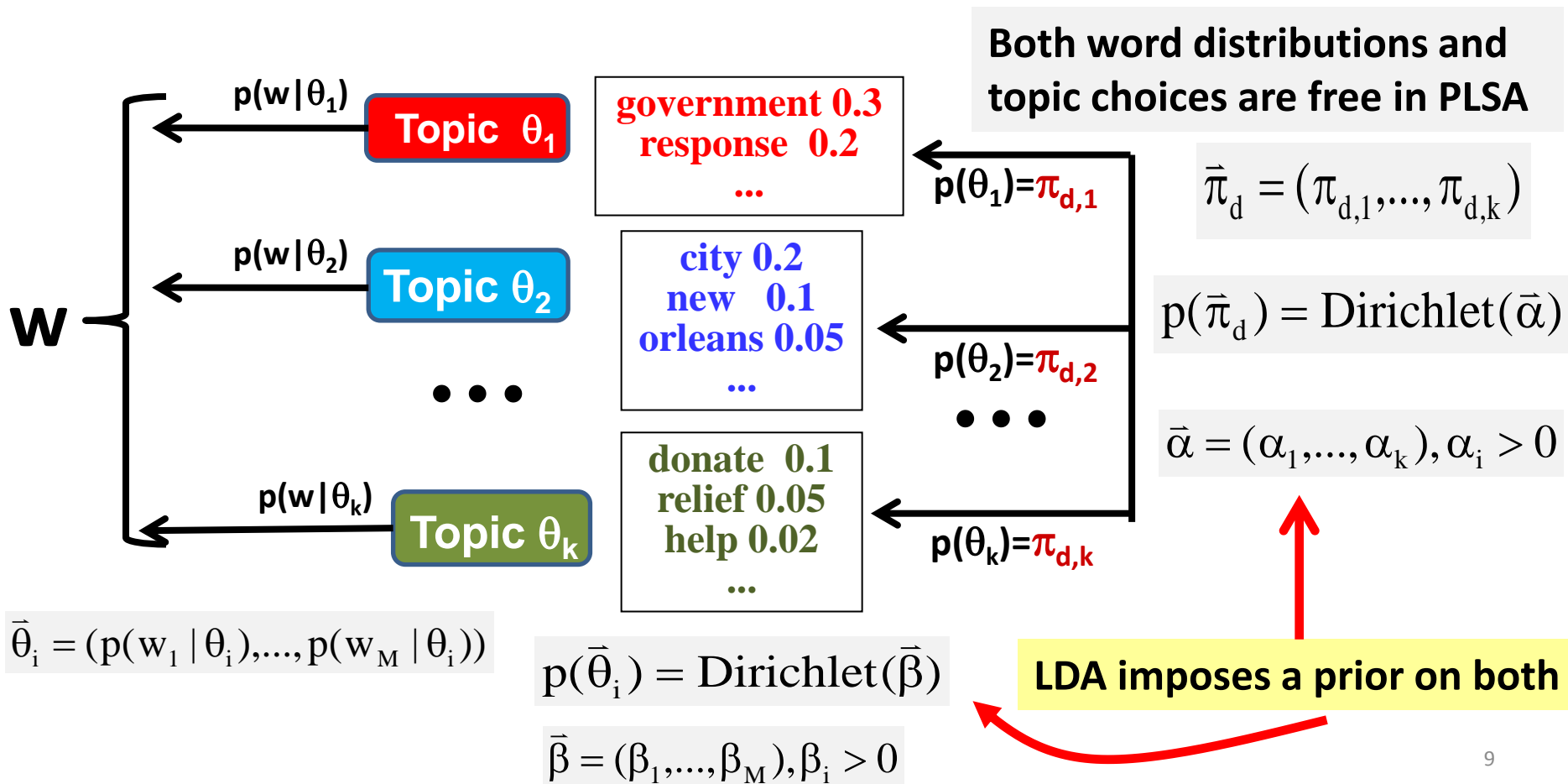**We may also set any parameter to a constant (including 0) as needed**

Deficiency of PLSA

- Not a generative model
  - Can't compute probability of a new document
  - Heuristic workaround is possible, though
- Many parameters ➔ high complexity of models
  - Many local maxima
  - Prone to overfitting
- Not necessarily a problem for text mining (only interested in fitting the "training" documents)

Latent Dirichlet Allocation (LDA)

- Make PLSA a generative model by imposing a Dirichlet prior on the model parameters ➔
  - LDA = Bayesian version of PLSA
  - Parameters are regularized
- Can achieve the same goal as PLSA for text mining purposes
  - Topic coverage and topic word distributions can be inferred using Bayesian inference

PLSA ➜ LDA

$$p(w|\theta_1)$$

Topic θ_1

government 0.3
response 0.2
...

$$p(w|\theta_2)$$

Topic θ_2

city 0.2
new 0.1
orleans 0.05
...

$$p(w|\theta_k)$$

Topic θ_k

donate 0.1
relief 0.05
help 0.02
...

W

$p(\theta_1)=\pi_{d,1}$

$p(\theta_2)=\pi_{d,2}$

$p(\theta_k)=\pi_{d,k}$

Both word distributions and topic choices are free in PLSA

$$\vec{\pi}_d = (\pi_{d,1},...,\pi_{d,k})$$

$$p(\vec{\pi}_d) = \text{Dirichlet}(\vec{\alpha})$$

$$\vec{\alpha} = (\alpha_1,...,\alpha_k), \alpha_i > 0$$

$$\vec{\theta}_i = (p(w_1 | \theta_i),...,p(w_M | \theta_i))$$

$$p(\vec{\theta}_i) = \text{Dirichlet}(\vec{\beta})$$

LDA imposes a prior on both

$$\vec{\beta} = (\beta_1,...,\beta_M), \beta_i > 0$$

Likelihood Functions for PLSA vs. LDA

PLSA

$$p_d(w \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j)$$

$$\log p(d \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{w \in V} c(w,d) \log[\sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j)]$$

$$\log p(C \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{d \in C} \log p(d \mid \{\theta_j\}, \{\pi_{d,j}\})$$

Core assumption in all topic models

LDA

$$p_d(w \mid \{\theta_j\}, \{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j)$$

$$\log p(d \mid \vec{\alpha}, \{\theta_j\}) = \int \sum_{w \in V} c(w,d) \log[\sum_{j=1}^{k} \pi_{d,j} p(w \mid \theta_j)] p(\vec{\pi}_d \mid \vec{\alpha}) d\vec{\pi}_d$$

$$\log p(C \mid \vec{\alpha}, \vec{\beta}) = \int \sum_{d \in C} \log p(d \mid \vec{\alpha}, \{\theta_j\}) \prod_{j=1}^{k} p(\theta_j \mid \vec{\beta}) d\theta_1 \ldots d\theta_k$$

PLSA component

Added by LDA

10

Parameter Estimation and Inferences in LDA

- Parameters can be estimated using ML estimator

$$(\hat{\vec{\alpha}}, \hat{\vec{\beta}}) = \arg\max_{\vec{\alpha}, \vec{\beta}} \log p(C \mid \vec{\alpha}, \vec{\beta})$$

How many parameters in LDA vs. PLSA?

- However, $\{\theta_j\}$ **and** $\{\pi_{d,j}\}$ must now be computed using posterior inference
  - Computationally intractable
  - Must resort to approximate inference
  - Many different inference methods are available

Summary of Probabilistic Topic Models

- Probabilistic topic models provide a general principled way of mining and analyzing topics in text with many applications
- Basic task setup:
  - Input: Text data
  - Output: k topics + proportions of these topics covered in each document
- PLSA is the basic topic model, often adequate for most applications
- LDA improves over PLSA by imposing priors
  - Theoretically more appealing
  - Practically, LDA and PLSA perform similarly for many tasks

Suggested Readings

- Blei, D. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. doi: 10.1145/2133806.2133826.

- Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. "Automatic Labeling of Multinomial Topic Models." *Proceedings of ACM KDD 2007*, pp. 490-499, DOI=10.1145/1281192.1281246.

- Yue Lu, Qiaozhu Mei, and Chengxiang Zhai. 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 14, 2 (April 2011), 178-203. DOI=10.1007/s10791-010-9141-9.