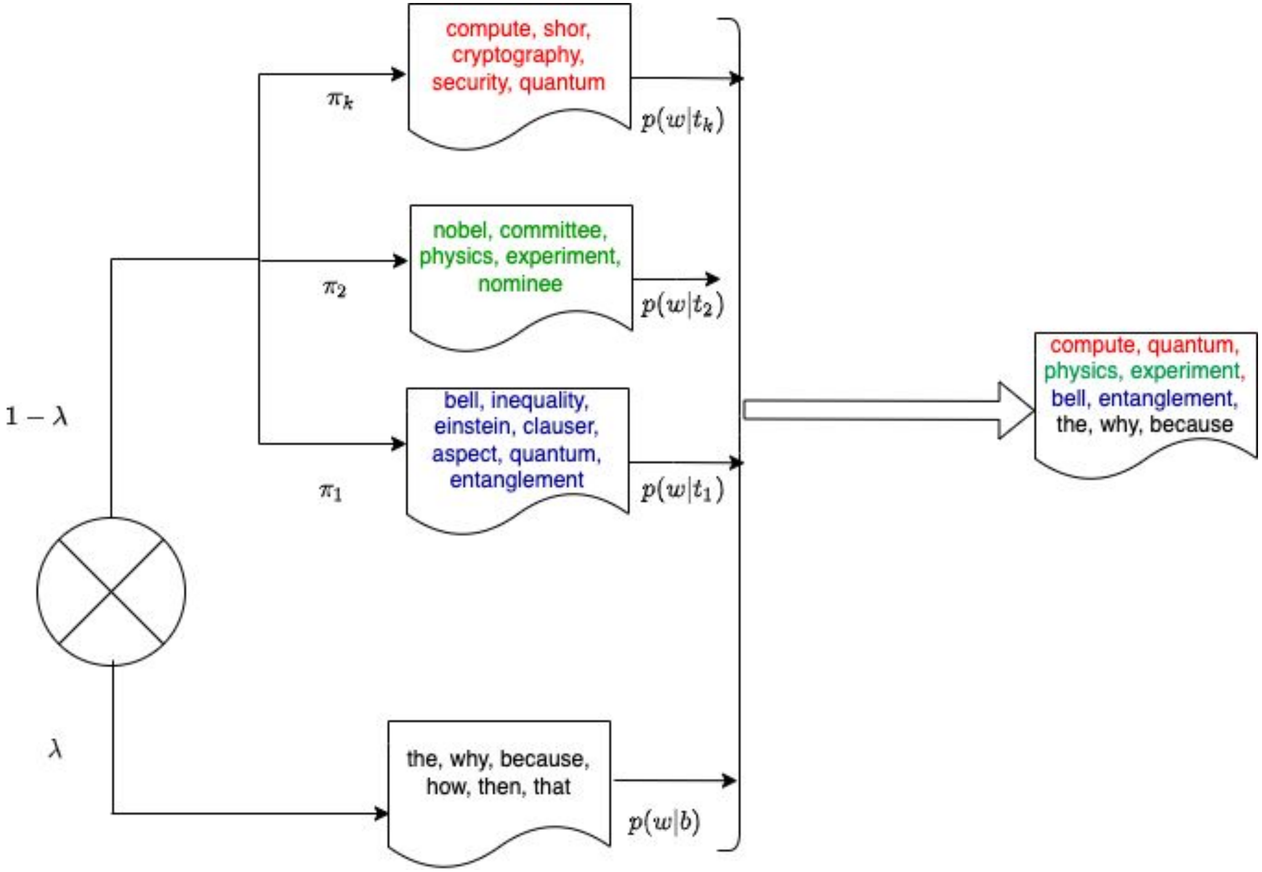
Topic Models

Lecture 2
Data Analysis
E0 259

Probabilistic Latent Semantic Analysis



Parameters to Estimate

- Given, document corpus D, number of topics k and vocabulary V, we need
- $\{\pi_{d,i}\}_{i=1,\ldots,k}$

Probability distribution of topics in each document

- $\{p(w|t_i)\}_{w \in V}$

Number of parameters to estimate is huge, function of corpus size and vocabulary size!!

Some Math

- Probability that a given word occurs in document
- $p(w) = \lambda p(w|b) + (1-\lambda)\sum_{i=1}^{k} \pi_{d,i} p(w|t_i)$
- $p(d) = \prod_{w \in V} p(w)^{c_w}$

Probability that document has all those words! We are still using Unigram language model

Product becomes Sum

Probability of word in document

- $log(p(d)) = \sum_{w \in V} c_w log(\lambda p(w|b) + (1-\lambda) \sum_{i=1}^{k} \pi_{d,i} p(w|t_i))$

- $p(D|\Theta) = \prod_{d \in D} p(d)$

Probability that document corpus occurs

- $log(p(D|\Theta)) = \sum_{d \in D} log(p(d))$

$$= \sum_{d \in D} \sum_{w \in W} c_w log(\lambda p(w|b) + (1-\lambda) \sum_{i=1}^{k} \pi_{d,i} p(w|t_i))$$

- $\arg\max_{\Theta} log(p(D|\Theta)) = \sum_{d \in D} log(p(d))$

$$= \sum_{d \in D} \sum_{w \in W} c_w log(\lambda p(w|b) + (1-\lambda) \sum_{i=1}^{k} \pi_{d,i} p(w|t_i))$$

- s.t. $\sum_{i=1}^{k} \pi_{d,k} = 1, \forall d \in D$ and $\sum_{w \in V} p(w|t_i) = 1, \forall i = 1, \ldots, k$

Use EM Algorithm to Solve - E-Step

- Same trick, introduce hidden variable $z_{d,w} \in b, 1, \ldots, k$

- $p^{n+1}(z_{d,w} = j) = \frac{\pi^n_{d,j} p^n(w|t_j)}{\sum_{i=1}^k \pi^n_{d,i} p^n(w|t_i)}$

Weighted average of word belonging to topic j

- $p^{n+1}(z_{d,w} = b) = \frac{\lambda p(w|b)}{\lambda p(w|b) + (1-\lambda)\sum_{i=1}^k \pi^n_{d,i} p^n(w|t_i)}$

Weighted average of word belonging to background topic

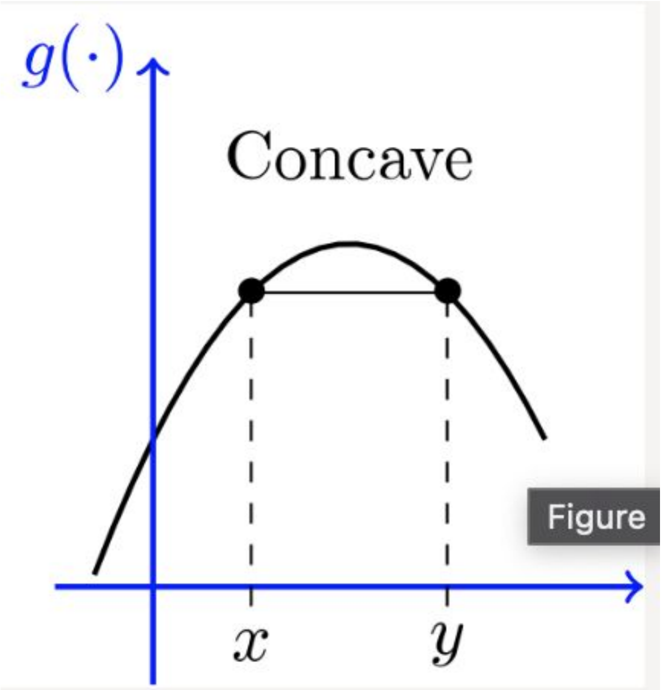
Use EM Algorithm to Solve - M-Step

- $\pi^{n+1}(d,j) = \frac{\sum_{w \in V} c_{w,d}(1-p^n(z_{d,w}=b))p^n(z_{d,w}=j)}{\sum_{i=1}^{k} \pi_{d,i}^{n} p^n(w|t_i)}$

- $p^{n+1}(w|t_j) = \frac{\sum_{d \in D} c_{w,d}(1-p^n(z_{d,w}=b))p(z_{d,w}=j)}{\sum_{w' \in V} \sum_{d \in D} c_{w',d}(1-p^n(z_{d,w'}=b))p^n(z_{d,w'}=j)}$

Jensen’s Inequality

For a concave function, $f(E[x]) \geq E[f(x)]$



EM Algorithm Solves a Lower Bound at Each Step

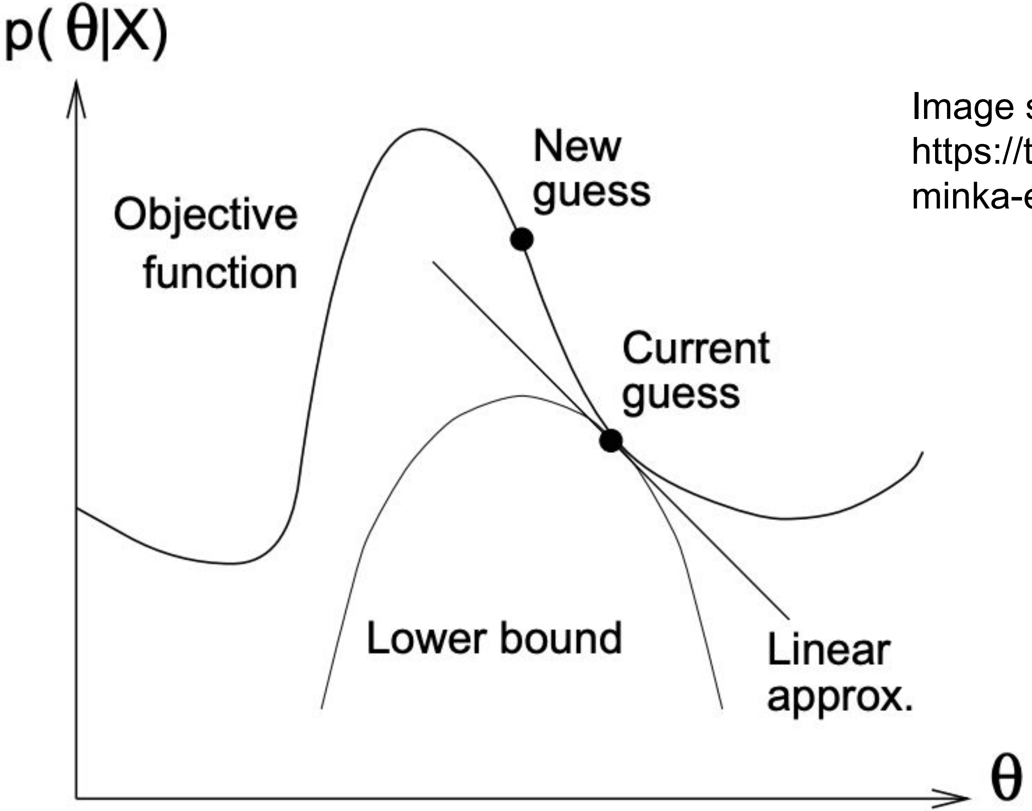


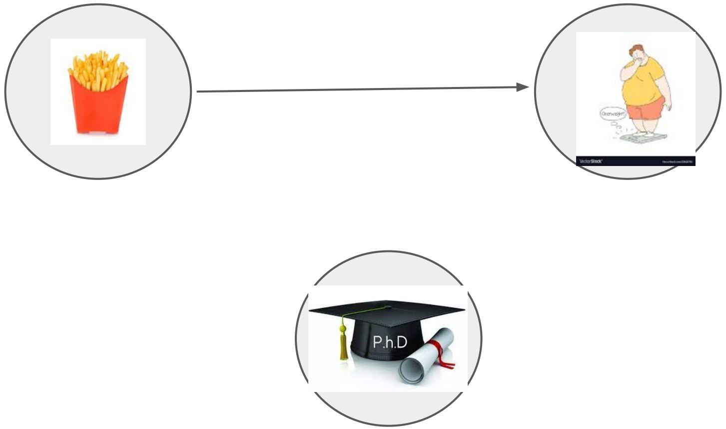
Image src:
https://tminka.github.io/papers/minka-em-tut.pdf

Challenges with PLSA

- Number of parameters = kV + kD
- Linear growth in parameters as documents and vocabulary increases
- Not generative model
    - If a new document comes, how do we know topic distribution for new document

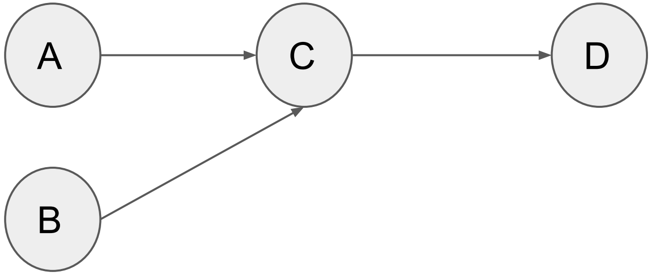
Graphical Models - Bayesian Networks

- Capture causality via nodes and edges
- P(Fries, Obesity, PhD) = P(Phd)P(Fries)P(Obesity | Fries)
- Captures conditional independence



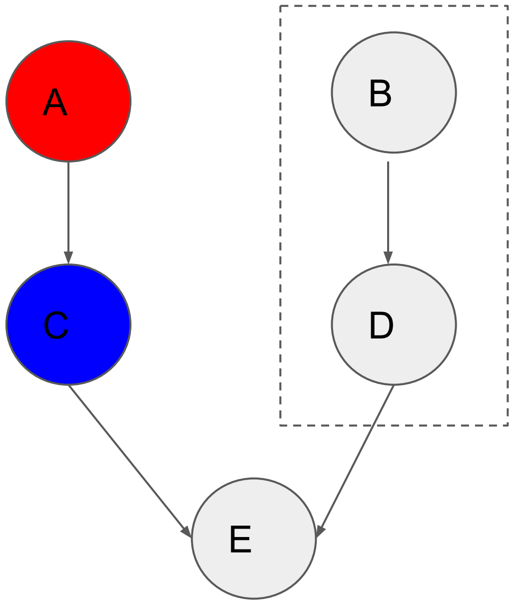
Bayesian Networks (contd.)

- Probability of occurrence of a node only dependent on it’s parents. P(X | parents(X)) e.g. P(C | A, B). P(D | C).
- Allows for efficient inference



Conditional Independence

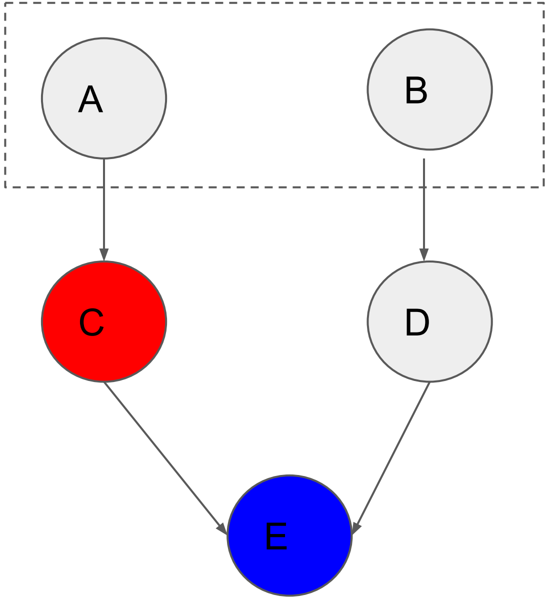
- Each node is conditionally independent of it's non descendents, given its parents.
- E.g. If A is known, then C independent of B and D.



Conditional Independence

- Each node is conditionally independent of it's non descendents, given its parents.

- E.g. If C is known, then E independent of A and B.

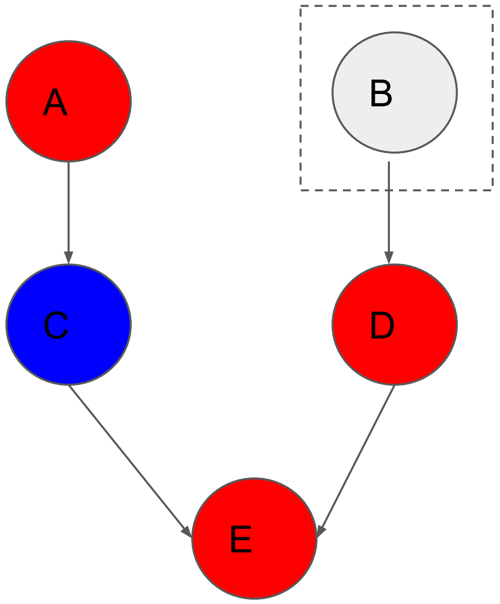


Conditional Independence

- Markov Blanket: Parents, Children and Children's parents

- Each node is independent of any other node, given it's Markov Blanket

- E.g. Given A, E and D, C is independent of B.



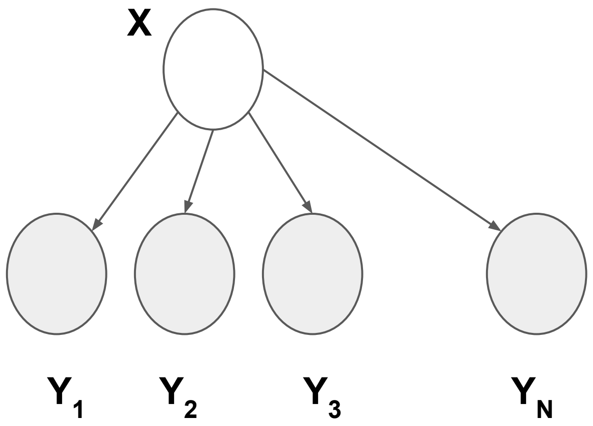
Graphical Models - Representation

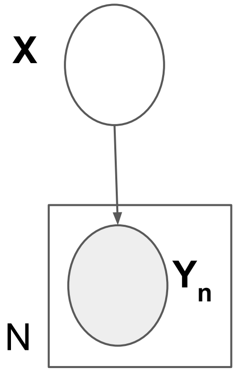
- Observed variables are shaded

- Plate denotes replicated structure

- In this graph

$$p(x, y_1, \ldots, y_n) = p(x) \prod_{i=1}^{N} p(y_i|x)$$



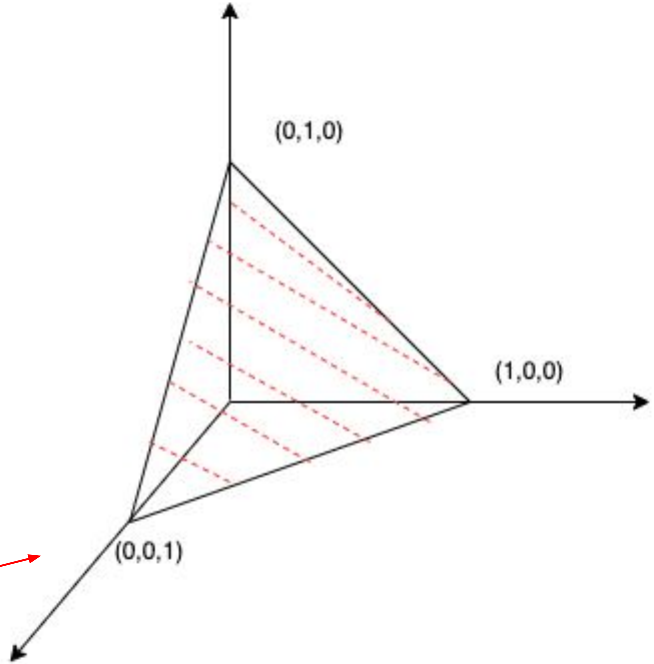


Dirichlet Distributions

- Used as a prior distribution in Bayesian statistics
- A distribution over n random variables

$$Dir(\alpha):$$
$$\mathbf{p}(\theta_\mathbf{1},\ldots,\theta_\mathbf{N}) = \frac{1}{\beta(\alpha)}\prod_{\mathbf{i=1}}^{\mathbf{N}}\theta_\mathbf{i}^{\alpha_\mathbf{i}-\mathbf{1}}\mathbf{I}\{\theta_\mathbf{i}\in\mathbf{S}\}$$

$$S = \{\theta_i \in \mathbb{R}, \theta_i \geq 0, \sum_{i=1}^{N}\theta_i = 1\}$$



Probability Simplex

Dirichlet Distributions (contd.)

- No Γ is not a function of some Δ, it is the Γ function:)
- $$\beta(\alpha) = \frac{\mathbf{\Gamma}(\alpha_{\mathbf{0}})}{\prod_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{\Gamma}(\alpha_{\mathbf{i}})}$$

$$\alpha_0 = \sum_{i=1}^{N} \alpha_i$$

$$E[\theta_i] = \frac{\alpha_i}{\alpha_0}$$

$$\sigma^2(\theta_i) = \frac{\alpha_i(\alpha_0 - 1)}{\alpha_0^2(\alpha_0 + 1)}$$

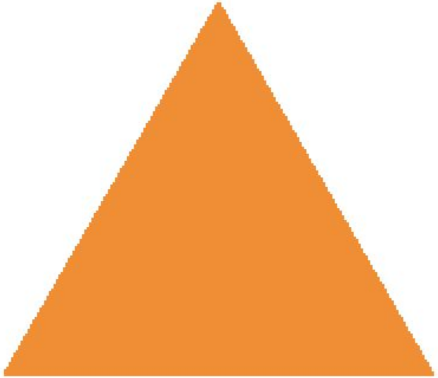
Properties of Dirichlet Distribution

$Dir(\alpha):$

$$\mathbf{p}(\theta_1, \ldots, \theta_N) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{N} \theta_i^{\alpha_i - 1} \mathbf{I}\{\theta_i \in \mathbf{S}\}$$

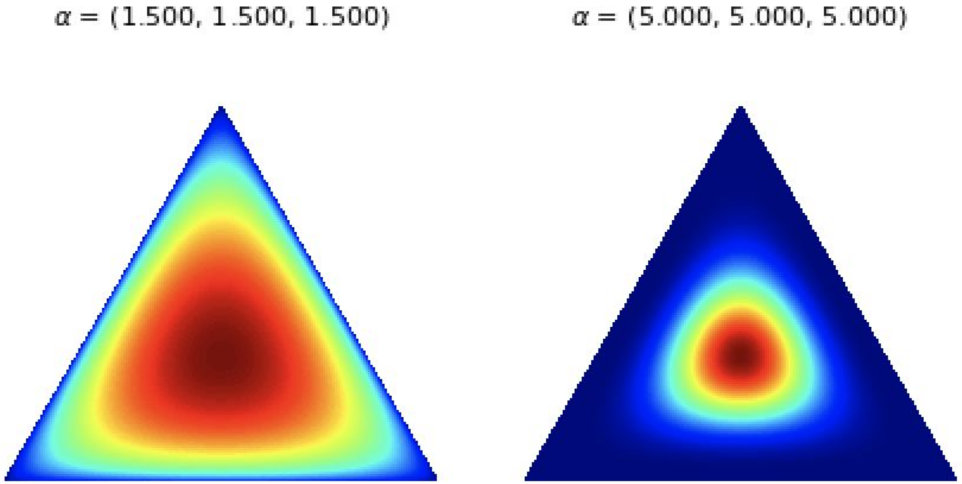
$$S = \{\theta_i \in \mathbb{R}, \theta_i \geq 0, \sum_{i=1}^{N} \theta_i = 1\}$$

α = (1.000, 1.000, 1.000)



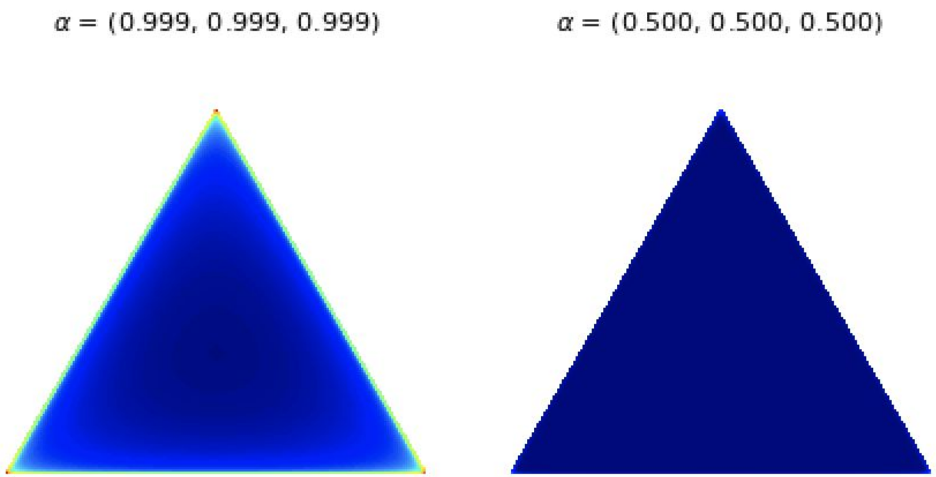
- When α=1, we get a uniform distribution over the simplex

Properties of Dirichlet Distribution (contd.)



- With equal α > 1, probability mass get more concentrated around center of simplex

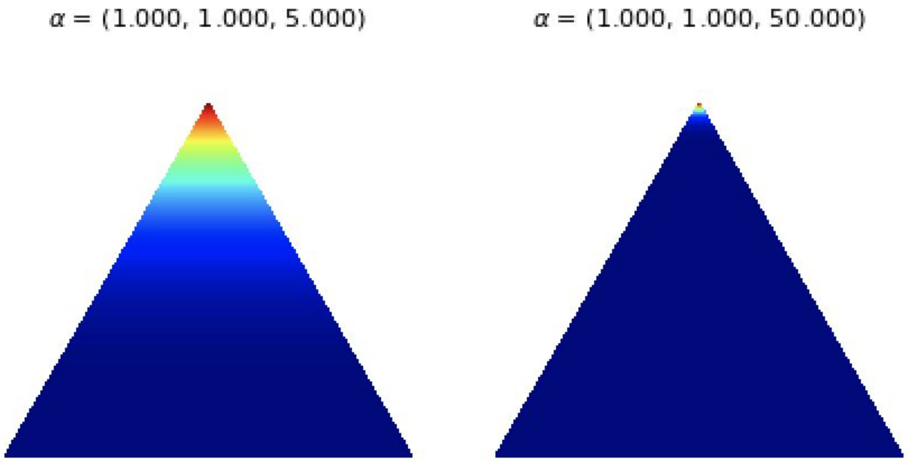
Properties of Dirichlet Distribution (contd.)



- With equal α < 1, probability mass get more concentrated around corners of simplex

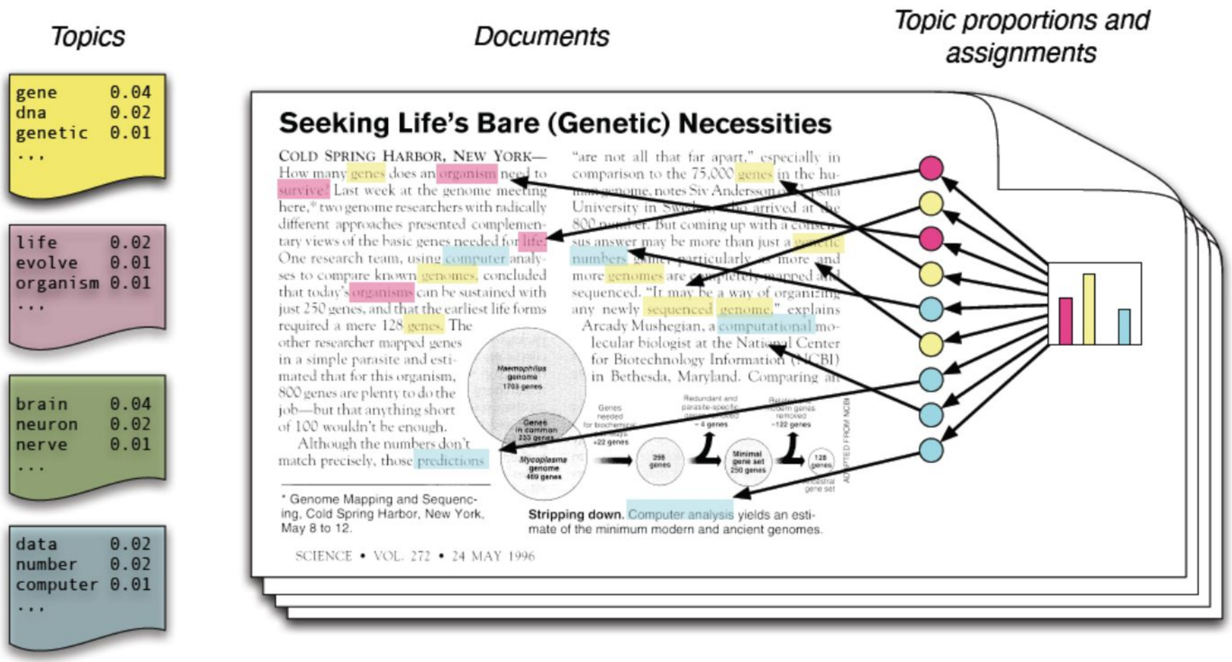
Properties of Dirichlet Distribution (contd.)

$$E[\theta_i] = \frac{\alpha_i}{\alpha_0}$$

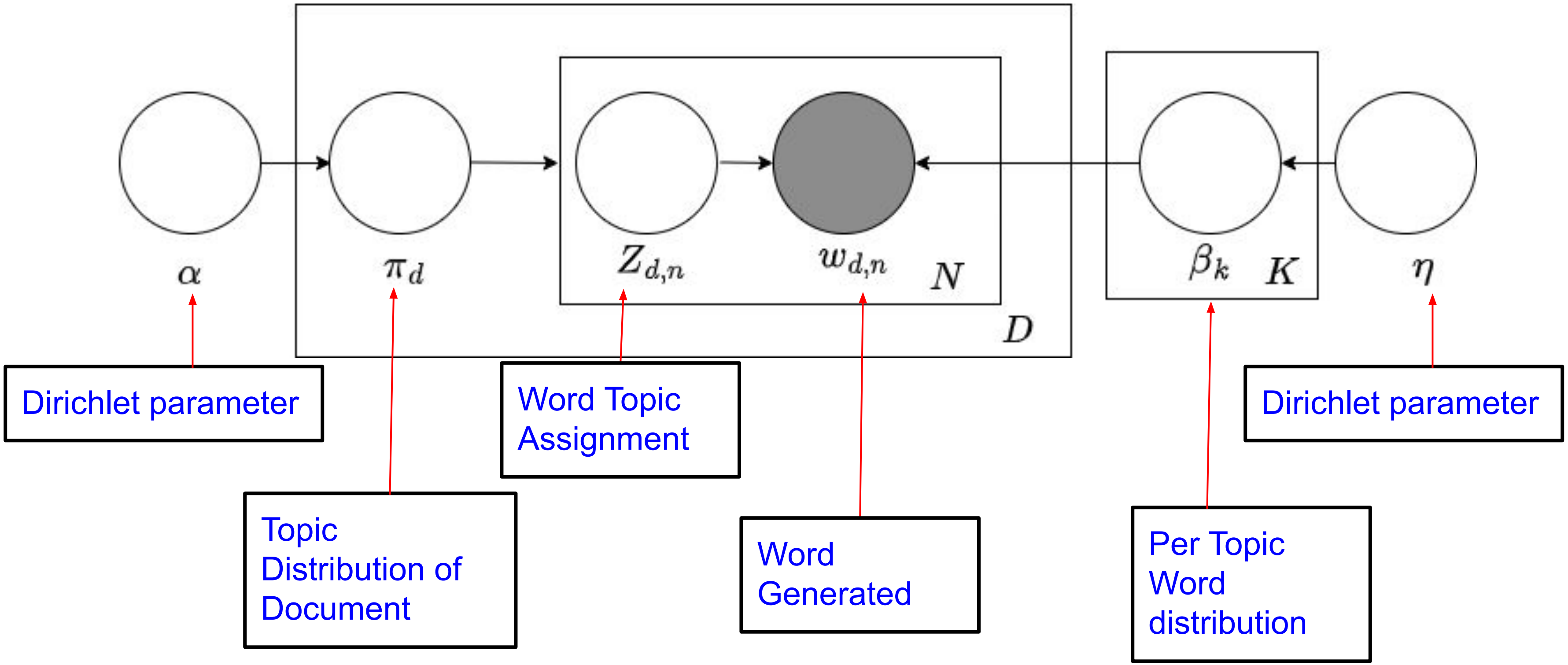


- With equal $\alpha > 1$ for one and rest remaining constant, probability mass get more concentrated around the large α

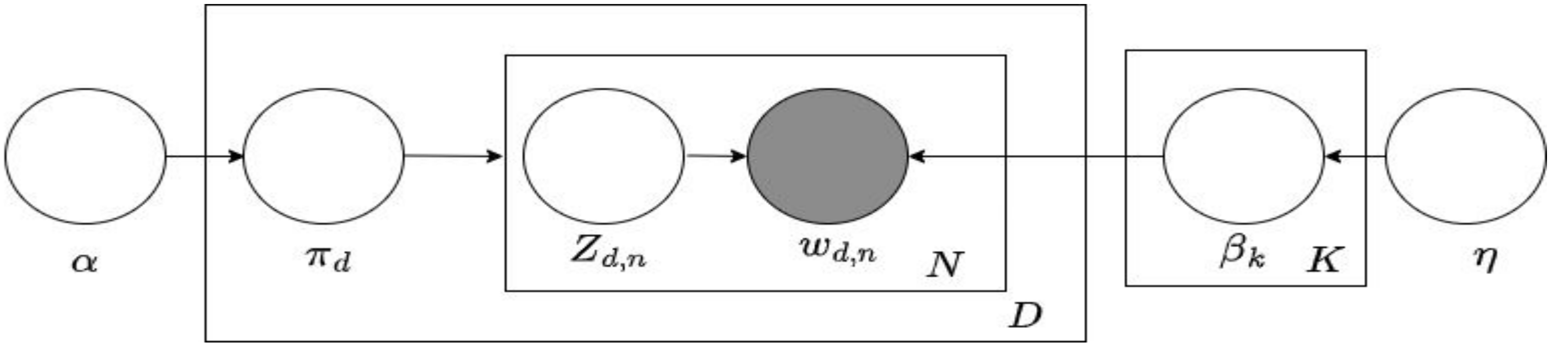
Latent Dirichlet Allocation



LDA Graphical Model



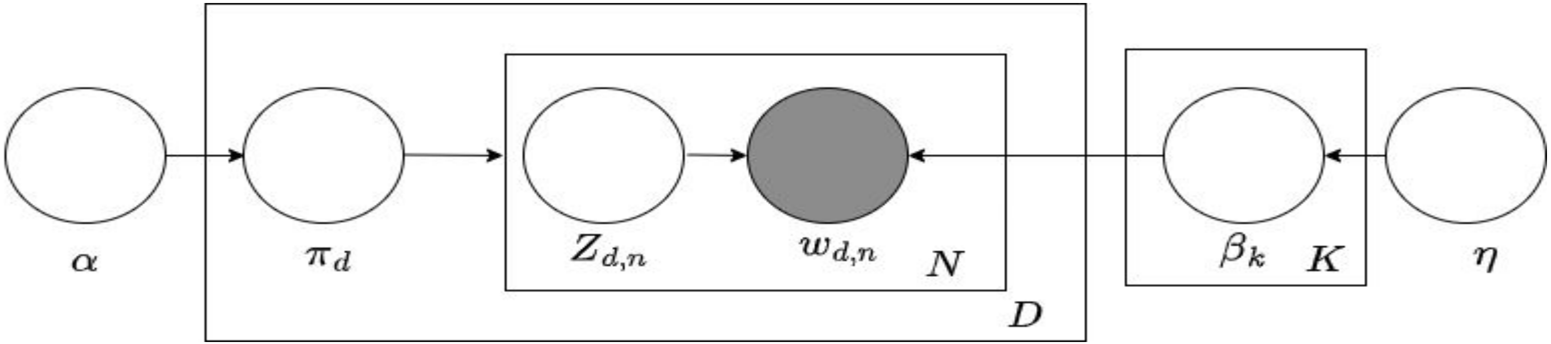
LDA Generative Model



- Draw each topic word distribution β from Dir(η)
- For each Document:
    - Topic distribution is π from Dir(α)
    - For each word:
        - Z is Mult(π)
        - W is Mult(β)

LDA Inference

- α is a hyper parameter
- We need to infer:
  - Per **word** topic assignment Z
  - Per **document** topic distribution π
  - Per **topic** word distribution β



Computing the Hidden Variable Distributions

$$p(\beta, \pi, \mathbf{Z}, \mathbf{W}) = \prod_{\mathbf{i=1}}^{\mathbf{K}} \mathbf{p}(\beta_{\mathbf{i}}) \prod_{\mathbf{i=1}}^{\mathbf{D}} \mathbf{p}(\pi_{\mathbf{d}}) \left(\prod_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{p}(\mathbf{Z}_{\mathbf{d,n}}|\pi_{\mathbf{d}})\mathbf{p}(\mathbf{w}_{\mathbf{d,n}}|\beta, \mathbf{z}_{\mathbf{d,n}})\right)$$

Joint probability distribution from graphical model

$$p(\beta, \pi, \mathbf{Z}|\mathbf{W}) = \frac{\mathbf{p}(\beta,\pi,\mathbf{Z},\mathbf{W})}{\mathbf{p}(\mathbf{W})}$$

$$p(\beta, \pi, \mathbf{Z}|\mathbf{W}) = \frac{\mathbf{p}(\beta,\pi,\mathbf{Z},\mathbf{W})}{\mathbf{p}(\mathbf{W})}$$

Posterior Distribution

Why This Model?

- In PLSA, essentially modeling each document in the training set comes from a point distribution over topics
- Hence for new unseen documents, there is no way to have a generative model
- LDA addresses this by having a generative model for the topic distribution of a document (essentially instead of a point, it is a distribution over the simplex).
- This gives it way more flexibility.
- Still the parameter space is large, how do we estimate it efficiently?