

Inference of Bayesian Network: MCMC

Jin Young Choi

Outline

- Latent Dirichlet Allocation (LDA) model(Topic Modelling)
 - Inference of LDA Model
 - Markov Chain Monte Carlo (MCMC)
 - Gibbs Sampling
 - Collapsed Gibbs Sampling for LDA
 - Estimation of Multinomial Parameters via Dirichlet Prior
-

LDA Model (Topic Modelling)

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

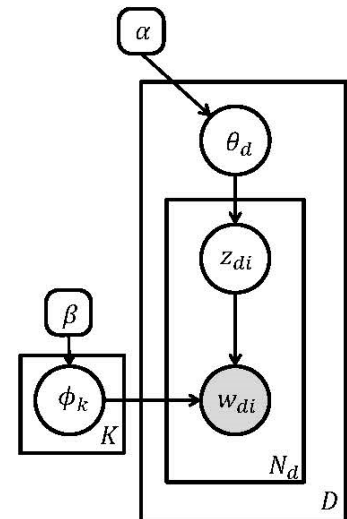
COLD SPRING HARBOR, NEW YORK— How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those **predictions** "are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a biologist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

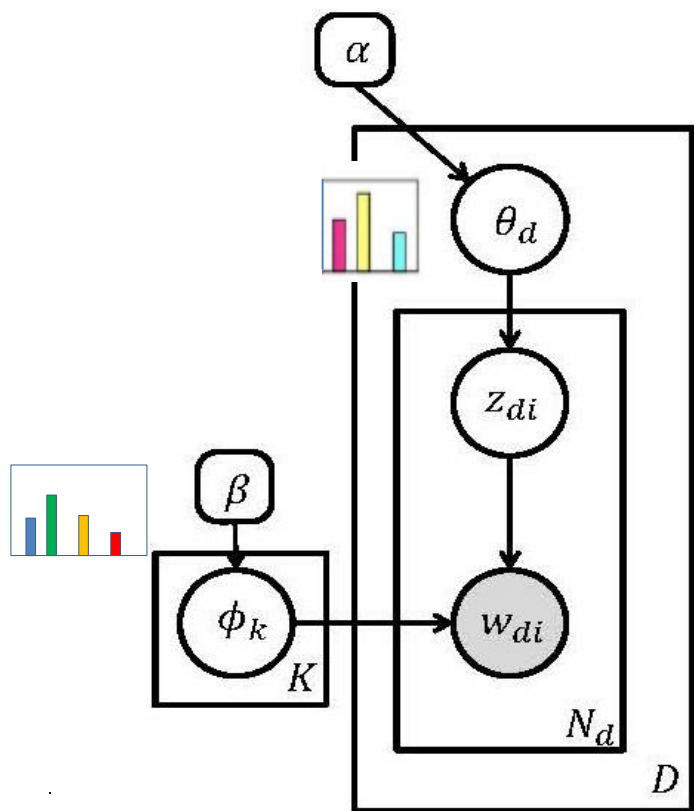


LDA Model

Likelihood: $p(w|z, \theta, \phi, \alpha, \beta)$

Posteriori: $p(z, \theta, \phi|w, \alpha, \beta)$

- $\text{Dir}(K, \alpha): p(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \mu_1^{\alpha_1-1} \dots \mu_K^{\alpha_K-1}$
- $\text{Mul}(K, \mu): p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$



Notations

D : the number of documents.

N_d : the number of words in d -th document.

K : the number of topics.

α : Dirichlet prior on the per-document topic distributions.

β : Dirichlet prior on the per-topic word distribution.

θ_d : topic distribution for d -th document.

ϕ_k : word distribution for topic k .

z_{di} : the topic for the i -th word in d -th document.

w_{di} : the specific word.

$$\{w_{d1}, w_{d2}, \dots, w_{dN_d}\}$$

Mathematical description

Choose $\theta_d \sim \text{Dir}(\alpha)$.

Choose $\phi_k \sim \text{Dir}(\beta)$.

Choose a topic $z_{ji} | \theta_d \sim \text{Multi}(\theta_d)$.

Choose a word $w_{ji} | \phi_k, z_{di} \sim \text{Multi}(\phi_{z_{di}})$.

LDA Model

- Dir(K, α): $p(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \mu_1^{\alpha_1-1} \dots \mu_K^{\alpha_K-1}$
- Mul(K, μ): $p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$

$$p(\phi, \theta, z, w | \alpha, \beta) = \left(\prod_{k=1}^K p(\phi_k | \beta) \right) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{i=1}^{N_d} p(z_{di} | \theta_d) p(w_{di} | z_{di}, \phi)$$

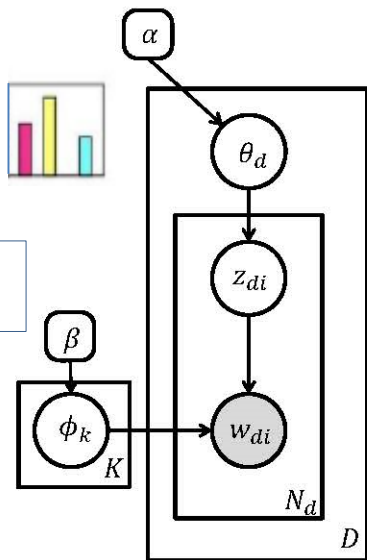
$$p(\phi_1, \phi_2, \dots, \phi_K | \beta) = \prod_{k=1}^K p(\phi_k | \beta), \quad \phi_k | \beta \sim \text{Dir}(\phi_k | \beta).$$

$$p(\theta_1, \dots, \theta_D | \alpha) = \prod_{d=1}^D p(\theta_d | \alpha) \quad \theta_d | \alpha \sim \text{Dir}(\theta_d | \alpha).$$

$$p(z_{d1}, z_{d2}, \dots, z_{dN_d} | \theta_d) = \prod_{i=1}^{N_d} p(z_{di} | \theta_d),$$

$$z_{di} | \theta_d \sim \text{Multi}(z_{di} | \theta_d),$$

$$w_{di} | z_{di}, \phi_1, \phi_2, \dots, \phi_K \sim \text{Multi}(w_{di} | \phi_{z_{di}}).$$



Inference of LDA Model

Likelihood: $p(w|z, \theta, \phi, \alpha, \beta)$

Posteriori: $p(z, \theta, \phi|w, \alpha, \beta)$

- Maximum A posteriori Probability (MAP) given observation w , α, β

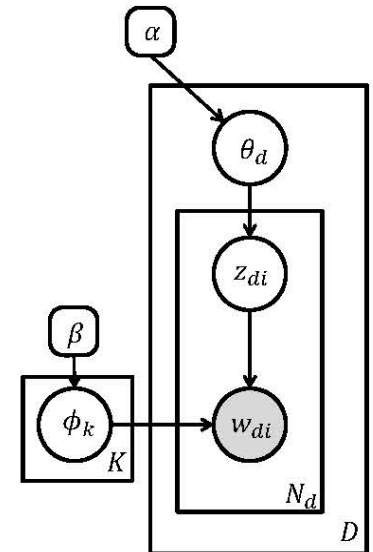
$$\hat{\phi}, \hat{\theta}, \hat{z} = \arg \max_{\phi, \theta, z} p(\phi, \theta, z|w, \alpha, \beta),$$

Not Convex

Closed-form solution is not available

$$\begin{aligned} p(\phi, \theta, z|w, \alpha, \beta) &= \frac{p(\phi, \theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}, \\ &= \frac{p(\phi, \theta, z, w|\alpha, \beta)}{\int_{\phi} \int_{\theta} \sum_z p(\phi, \theta, z, w|\alpha, \beta) d\theta d\phi}. \end{aligned}$$

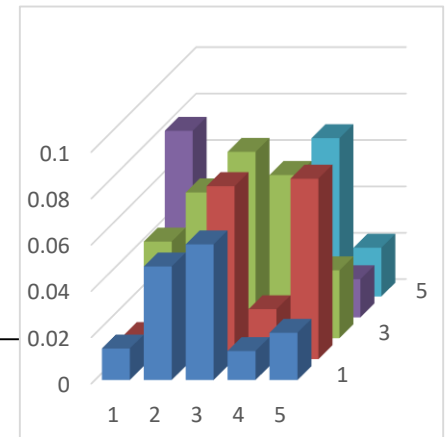
Likelihoods



$$p(\phi, \theta, z, w|\alpha, \beta) = \left(\prod_{k=1}^K p(\phi_k|\beta) \right) \prod_{d=1}^D p(\theta_d|\alpha) \prod_{i=1}^{N_d} p(z_{di}|\theta_d) p(w_{di}|z_{di}, \phi).$$

Discrete Variables, Dirichlet

- Parameters: $\alpha_1, \dots, \alpha_K > 0$ (concentration hyper-parameter)
- Support: $\mu_1, \dots, \mu_K \in (0,1)$ where $\sum_{i=1}^K \mu_i = 1$
- Dir(K, α): $p(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \mu_1^{\alpha_1-1} \dots \mu_K^{\alpha_K-1}$
- Mul(K, μ): $p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$
- Dir($K, c + \alpha$): $p(\mu | x, \alpha) \propto p(x | \mu) p(\mu | \alpha)$
 where $c = (c_1, \dots, c_K)$ is number of occurrences
- $E[\mu_k] = \frac{c_k + \alpha_k}{\sum_{i=1}^K (c_i + \alpha_i)}$



Markov Chain Monte Carlo (MCMC)

- Markov Chain Monte Carlo (MCMC) framework

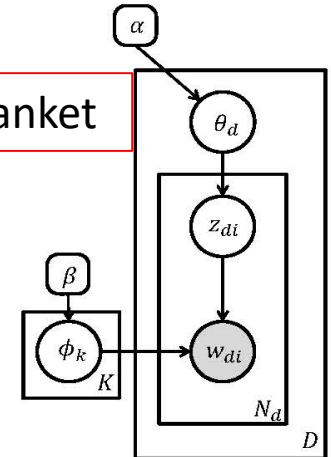
Posteriors

$$p(\theta_d | z, \alpha) = \frac{\overbrace{p(z | \theta_d)}^{\text{Multinomial}} \overbrace{p(\theta_d | \alpha)}^{\text{Dirichlet}}}{p(z | \alpha)}$$

$$= \text{Dir}(\theta_d | h_{\theta}(d, \cdot) + \alpha), \quad h_{\theta}(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

$$p(\phi_k | z, w, \beta) = \text{Dir}(\phi_k | h_{\phi}(k, \cdot) + \beta). \quad h_{\phi}(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]$$

Markov Blanket



$$\hat{\theta}_d(k) = E[\theta_d(k) | h_{\theta}(d, \cdot) + \alpha] = \frac{h_{\theta}(d, k) + \alpha(k)}{\sum_{k=1}^K [h_{\theta}(d, k) + \alpha(k)]}$$

$$\hat{\phi}_k(v) = E[\phi_k(v) | h_{\phi}(k, \cdot) + \beta] = \frac{h_{\phi}(k, v) + \beta(v)}{\sum_{v=1}^V [h_{\phi}(k, v) + \beta(v)]}$$

i: 1, 2, 3, 4, 5, 6, 7, 8, 9
w: 1, 3, 2, 3, 3, 5, 4, 1, 6
z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 $h_{\theta}(d, 2): 5$

i: 1, 2, 3, 4, 5, 6, 7, 8, 9
w: 1, 3, 2, 3, 3, 5, 4, 1, 6
z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 $h_{\phi}(1, 3): 1$
 $h_{\phi}(2, 3): 2$

Markov Chain Monte Carlo (MCMC)

- Markov Chain Monte Carlo (MCMC) framework

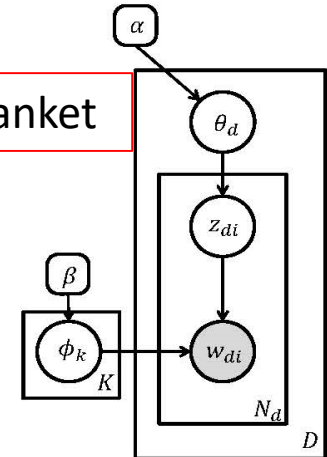
Posteriors

$$p(\theta_d | z, \alpha) = \frac{\overbrace{p(z|\theta_d)}^{\text{Multinomial}} \overbrace{p(\theta_d|\alpha)}^{\text{Dirichlet}}}{p(z|\alpha)}$$

$$= \text{Dir}(\theta_d | h_{\theta}(d, \cdot) + \alpha), \quad h_{\theta}(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

$$p(\phi_k | z, w, \beta) = \text{Dir}(\phi_k | h_{\phi}(k, \cdot) + \beta). \quad h_{\phi}(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]$$

Markov Blanket



Gibbs sampling

$p(z|w, \phi_k, \alpha)$

$p(z|w, \phi_k, \alpha)$

$$\hat{\theta}_d(k) = E[\theta_d(k) | h_{\theta}(d, \cdot) + \alpha] = \frac{h_{\theta}(d, k) + \alpha(k)}{\sum_{k=1}^K [h_{\theta}(d, k) + \alpha(k)]}$$

$$\hat{\phi}_k(v) = E[\phi_k(v) | h_{\phi}(k, \cdot) + \beta] = \frac{h_{\phi}(k, v) + \beta(v)}{\sum_{v=1}^V [h_{\phi}(k, v) + \beta(v)]}$$

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 $h_{\theta}(d,2): 2$

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 $h_{\phi}(1,3): 1$
 $h_{\phi}(2,3): 2$

Gibbs Sampling (Review)

$$p(z) = p(z_1, z_2, \dots, z_N)$$

1. Randomly initialize each $z_i^1 \in \{1, 2, \dots, K\}$, where $i = 1, 2, \dots, N$,
 2. For each step $t = 1, 2, \dots, T$:
 - Replace z_1^t by a new value z_1^{t+1} , sampling $z_1^{t+1} \sim p(z_1 | z_2^t, z_3^t, \dots, z_N^t)$.
 - Replace z_2^t by a new value z_2^{t+1} , sampling $z_2^{t+1} \sim p(z_2 | z_1^{t+1}, z_3^t, \dots, z_N^t)$.
 - ...
 - Replace z_j^t by a new value z_j^{t+1} ,
sampling $z_j^{t+1} \sim p(z_j | z_1^{t+1}, \dots, z_{j-1}^{t+1}, z_{j+1}^t, \dots, z_N^t)$.
 - Replace z_N^t by a new value z_N^{t+1} , sampling $z_N^{t+1} \sim p(z_N | z_1^{t+1}, \dots, z_{N-1}^{t+1})$.
-

Collapsed Gibbs Sampling for LDA

- Latent Variables

- θ : topic distribution in a document
- ϕ : word distribution in a topic
- z : topic assignment to a word w
- $p(\theta, \phi, z|w, \alpha, \beta)$

$$\begin{aligned}
 p(\theta_d|z, \alpha) &= \frac{\overbrace{p(z|\theta_d)}^{\text{Multinomial Dirichlet}} \overbrace{p(\theta_d|\alpha)}^{\text{Multinomial Dirichlet}}}{p(z|\alpha)} \\
 &= \text{Dir}(\theta_d|h_\theta(d, \cdot) + \alpha), \\
 p(\phi_k|z, w, \beta) &= \text{Dir}(\phi_k|h_\phi(k, \cdot) + \beta).
 \end{aligned}$$

$$p(z|w, \alpha, \beta)$$

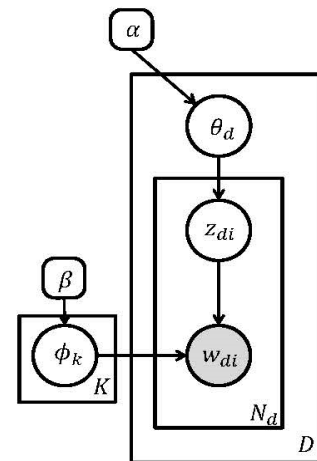
$$p(z|w, \phi_k, \alpha)$$

Collapsed Gibbs Sampling

- Collapsed Gibbs Sampling

- θ and ϕ are induced by the association between z and w
- z is sufficient statistic to estimate θ and ϕ
- Simpler algorithm can be used by sampling only z after marginalizing θ and ϕ .

- $p(z|w, \alpha, \beta) = \iint p(\theta, \phi, z|w, \alpha, \beta) d\theta d\phi$



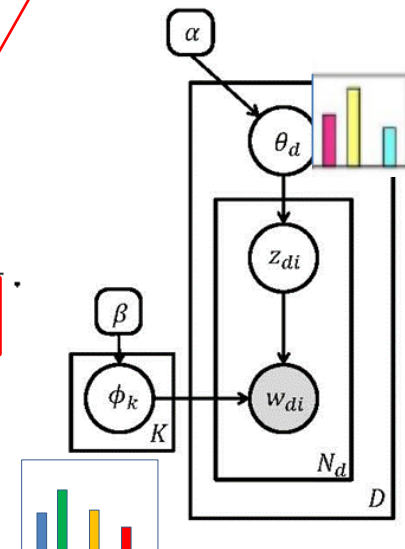
Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA (Topic Labelling, Unsupervised Learning)

$$p(z|w, \alpha, \beta)$$

$$\begin{aligned}
 p(z_{di}|z_{-di}, w, \alpha, \beta) &= \frac{p(z_{di}, z_{-di}, w|\alpha, \beta)}{p(z_{-di}, w|\alpha, \beta)} \\
 &= \frac{p(z, w|\alpha, \beta)}{p(z_{-di}, w|\alpha, \beta)} \\
 &= \frac{p(z|\alpha, \beta)p(w|z, \alpha, \beta)}{p(z_{-di}|\alpha, \beta)p(w_{di}, w_{-di}|z_{-di}, \alpha, \beta)} \\
 &= \frac{p(z|\alpha)p(w|z, \beta)}{p(z_{-di}|\alpha)p(w_{-di}|z_{-di}, \beta)p(w_{di}|\alpha, \beta)}
 \end{aligned}$$

Not related with z_{-di}



Collapsed Gibbs Sampling for LDA

- Dirichlet and multinomial probability into $p(\phi|\beta)$ and $p(w|z, \phi)$

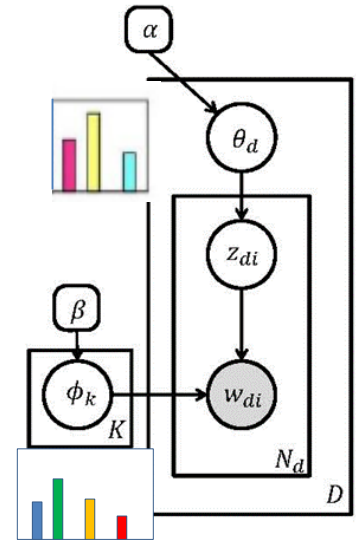
$$\begin{aligned}
 p(w|z, \beta) &= \int p(\phi|\beta)p(w|z, \phi)d\phi \\
 &= \int \left\{ \prod_{k=1}^K p(\phi_k|\beta) \right\} \prod_{d=1}^D \prod_{i=1}^{N_d} p(w_{di}|z_{di}, \phi) d\phi \\
 &= \int \left\{ \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_k(v)^{\beta(v)-1} \right\} \prod_{d=1}^D \prod_{i=1}^{N_d} \phi_{z_{di}}(w_{di}) d\phi,
 \end{aligned}$$

where

$$B(\beta) = \frac{\prod_{v=1}^V \Gamma(\beta(v))}{\Gamma\left(\sum_{v=1}^V \beta(v)\right)}$$

- Dir(K, α): $p(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \mu_1^{\alpha_1-1} \dots \mu_K^{\alpha_K-1}$

- Mul(K, μ): $p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$



Collapsed Gibbs Sampling for LDA

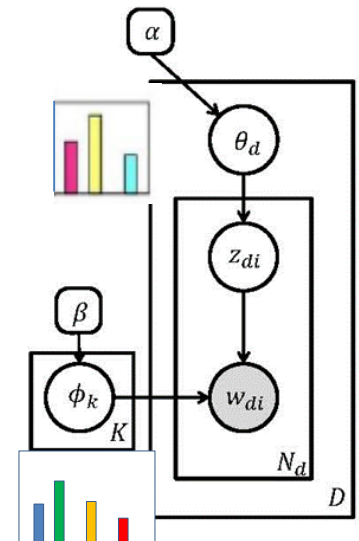
- The number of times that the word $w_{di} = v$ is assigned to the topic $z_{di} = k$

$$\prod_{d=1}^D \prod_{i=1}^{N_d} \phi_{z_{di}}(w_{di}) = \prod_{k=1}^K \prod_{v=1}^V \{\phi_k(v)\}^{h(k,v)}$$

where $h_\phi(k, v) \in N^{K \times V}$ denotes the histogram matrix which counts the number of times given by

$$h_\phi(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k],$$

w: 1, 2, 2, 3, 3, 5, 4, 1, 6
z: 1, 2, 2, 1, 1, 1, 2, 1, 2
h(2,3): 0
h(1,3): 2



Collapsed Gibbs Sampling for LDA

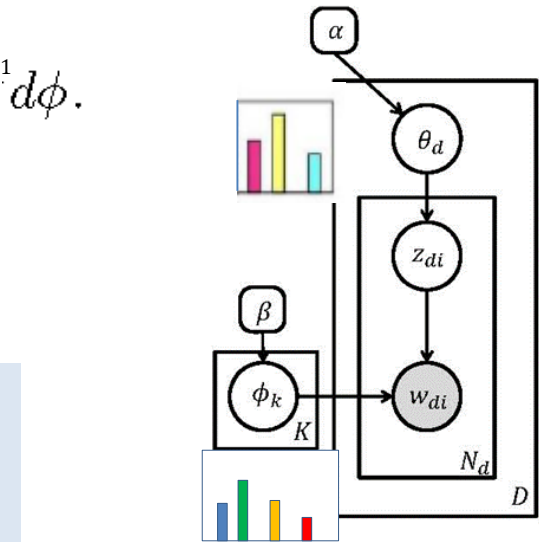
- In consequence

$$\begin{aligned}
 p(w|z, \beta) &= \int \left\{ \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_k(v)^{\beta(v)-1} \right\} \prod_{d=1}^D \prod_{i=1}^{N_d} \phi_{z_{di}}(w_{di}) d\phi \\
 &= \int \left\{ \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_k(v)^{\beta(v)-1} \right\} \prod_{k=1}^K \prod_{v=1}^V \{\phi_k(v)\}^{h_{\phi}(k,v)} d\phi \\
 &= \prod_{k=1}^K \frac{1}{B(\beta)} \int \prod_{v=1}^V \{\phi_k(v)\}^{h_{\phi}(k,v) + \beta(v) - 1} d\phi.
 \end{aligned}$$

$B(h_{\phi}(k, \cdot) + \beta)$

- Dir(K, α): $p(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \mu_1^{\alpha_1-1} \dots \mu_K^{\alpha_K-1}$

- Mul(K, μ): $p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$



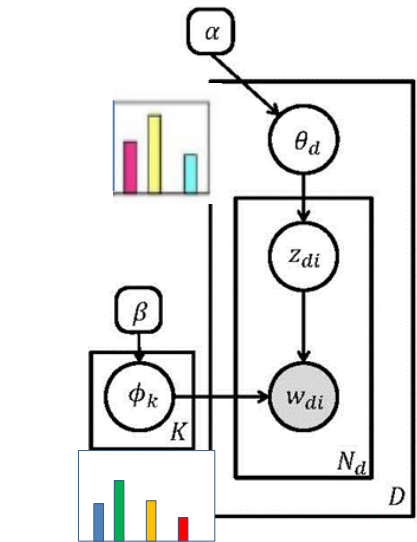
Collapsed Gibbs Sampling for LDA

- Integrating PDF

$$\begin{aligned}
 p(w|z, \beta) &= \prod_{k=1}^K \frac{1}{B(\beta)} \int \prod_{v=1}^V \{\phi_k(v)\}^{h_{\phi}(k,v) + \beta(v) - 1} d\phi \\
 &= \prod_{k=1}^K \frac{B(h_{\phi}(k, \cdot) + \beta)}{B(\beta)} \underbrace{\int \frac{1}{B(h_{\phi}(k, \cdot) + \beta)} \prod_{v=1}^V \{\phi_k(v)\}^{h_{\phi}(k,v) + \beta(v) - 1} d\phi}_{=1 \text{ (Integral of pdf)}}
 \end{aligned}$$

$$= \prod_{k=1}^K \frac{B(h_{\phi}(k, \cdot) + \beta)}{B(\beta)} \quad h_{\phi}(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k].$$

$w: 1, 2, 2, 3, 3, 5, 4, 1, 6$
 $z: 1, 2, 2, 1, 1, 1, 2, 1, 2$
 $h(2,3): 0$
 $h(1,3): 2$



$$p(z_{di} | z_{-di}, w, \alpha, \beta) = \frac{p(z | \alpha) p(w | z, \beta)}{p(z_{-di} | \alpha) p(w_{-di} | z_{-di}, \beta) p(w_{di} | \alpha, \beta)}.$$

Collapsed Gibbs Sampling for LDA

- In a similar manner

$$\begin{aligned}
 p(z|\alpha) &= \int p(\theta | \alpha) p(z|\theta) d\theta \\
 &= \int \prod_{d=1}^D p(\theta_d | \alpha) \prod_{i=1}^{N_d} p(z_{di} | \theta_d) d\theta \\
 &= \int \prod_{d=1}^D \left\{ \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_d(k)^{\alpha^{(k)}-1} \right\} \prod_{i=1}^{N_d} \theta_d(z_{di}) d\theta.
 \end{aligned}$$

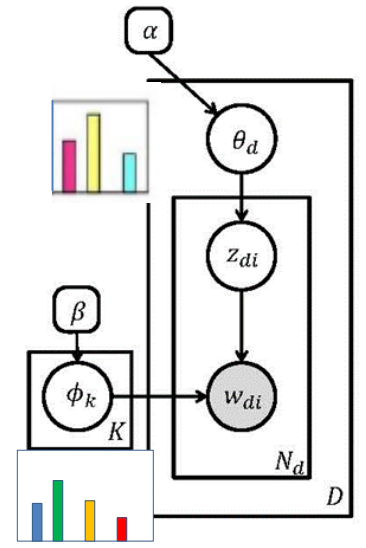
Topic portion corresponding to a word

$$\prod_{k=1}^K \theta_d(k)^{h_{\theta}(d,k)}$$

$$h_{\theta}(d,k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

$$\prod_{i=1}^{N_d} \theta_d(z_{di})$$

$w: 1, 2, 2, 3, 3, 5, 4, 1, 6$
 $z: 1, 2, 2, 1, 1, 1, 2, 1, 2$
 $h(d,2): 4$



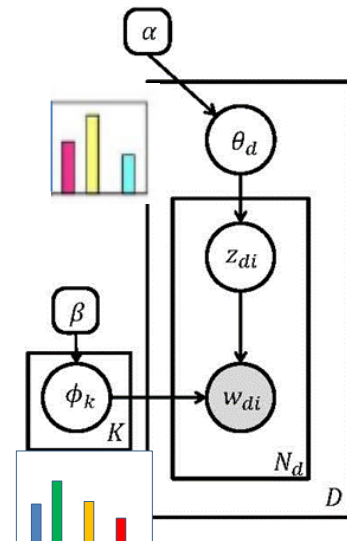
Collapsed Gibbs Sampling for LDA

- In a similar manner

$$\begin{aligned}
 p(z|\alpha) &= \int \prod_{d=1}^D \left\{ \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_d(k)^{\alpha^{(k)}-1} \right\} \prod_{k=1}^K \theta_d(k)^{h_{\theta}(d,k)} d\theta \\
 &= \int \prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_d(k)^{\alpha^{(k)}-1} \theta_d(k)^{h_{\theta}(d,k)} d\theta \\
 &= \int \prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_d(k)^{h_{\theta}(d,k) + \alpha^{(k)} - 1} d\theta \\
 &= \prod_{d=1}^D \frac{B(h_{\theta}(d, \cdot) + \alpha)}{B(\alpha)} \underbrace{\int \frac{1}{B(h_{\theta}(d, \cdot) + \alpha)} \prod_{k=1}^K \theta_d(k)^{h_{\theta}(d,k) + \alpha^{(k)} - 1} d\theta}_{=1 \text{ (Integral of pdf)}} \\
 &= \prod_{d=1}^D \frac{B(h_{\theta}(d, \cdot) + \alpha)}{B(\alpha)}.
 \end{aligned}$$

w: 1, 2, 2, 3, 3, 5, 4, 1, 6
z: 1, 2, 2, 1, 1, 1, 2, 1, 2
h(d,2): 4

$$h_{\theta}(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$



Collapsed Gibbs Sampling for LDA

- The joint distribution of words w and topic assignments z becomes

$$p(z, w | \alpha, \beta) = p(w | z, \beta) p(z | \alpha)$$

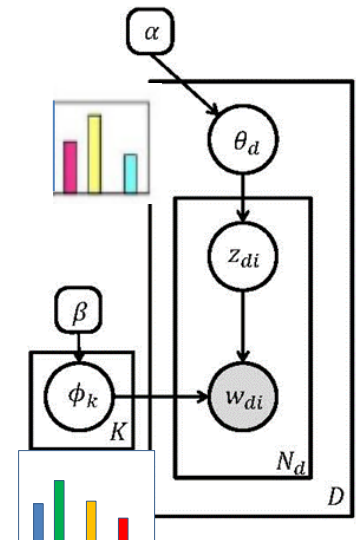
$$= \left\{ \prod_{k=1}^K \frac{B(h_\phi(k, \cdot) + \beta)}{B(\beta)} \right\} \left\{ \prod_{d=1}^D \frac{B(h_\theta(d, \cdot) + \alpha)}{B(\alpha)} \right\}$$

$$h_\phi(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]$$

$$h_\theta(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

w : 1, 3, 2, 3, 3, 5, 4, 1, 6
 z : 1, 2, 2, **2**, 1, 1, 2, 1, 2
 $h_\theta(d, 2)$: 5

w : 1, 3, 2, 3, 3, 5, 4, 1, 6
 z : 1, 2, 2, **2**, 1, 1, 2, 1, 2
 $h_\phi(1, 3)$: 1
 $h_\phi(2, 3)$: 2



Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA

$$p(z|w, \alpha, \beta)$$

$$\begin{aligned}
 p(z_{di}|z_{-di}, w, \alpha, \beta) &= \frac{p(z_{di}, z_{-di}, w|\alpha, \beta)}{p(z_{-di}, w|\alpha, \beta)} \\
 &= \frac{p(z, w|\alpha, \beta)}{p(z_{-di}, w|\alpha, \beta)} \\
 &= \frac{p(z|\alpha, \beta)p(w|z, \alpha, \beta)}{p(z_{-di}|\alpha, \beta)p(w_{di}, w_{-di}|z_{-di}, \alpha, \beta)} \\
 &= \frac{p(z|\alpha)p(w|z, \beta)}{p(z_{-di}|\alpha)p(w_{-di}|z_{-di}, \beta) \color{red}{p(w_{di}|\alpha, \beta)}}
 \end{aligned}$$

$$\begin{aligned}
 &p(w|z, \beta)p(z|\alpha) \\
 &= \left\{ \prod_{k=1}^K \frac{B(h_\phi(k, \cdot) + \beta)}{B(\beta)} \right\} \left\{ \prod_{d=1}^D \frac{B(h_\theta(d, \cdot) + \alpha)}{B(\alpha)} \right\}
 \end{aligned}$$

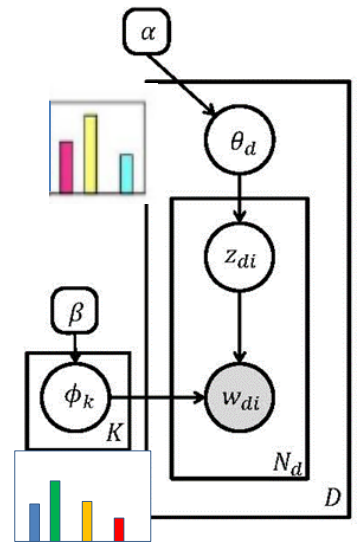
$$h_\phi(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]$$

$$h_\theta(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 h_θ(d,2): 5

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 h_φ(1,3): 1
 h_φ(2,3): 2

Not related with z_{-di}



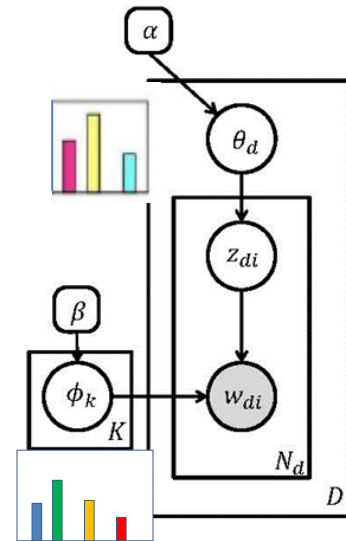
Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA

$$p(z_{di} | z_{-di}, w, \alpha, \beta) \propto \frac{p(z|\alpha)p(w|z, \beta)}{p(z_{-di}|\alpha)p(w_{-di}|z_{-di}, \beta)\cancel{p(w_{di}|\alpha, \beta)}}$$

$$p(z_{di} | z_{-di}, w, \alpha, \beta) \propto \frac{\left\{ \prod_{k=1}^K \frac{B(h_\phi(k, \cdot) + \beta)}{B(\beta)} \right\} \left\{ \prod_{d=1}^D \frac{B(h_\theta(d, \cdot) + \alpha)}{B(\alpha)} \right\}}{\left\{ \prod_{k=1}^K \frac{B(h_\phi(k, -di) + \beta)}{B(\beta)} \right\} \left\{ \prod_{d=1}^D \frac{B(h_\theta(d, -di) + \alpha)}{B(\alpha)} \right\}}$$

$$= \prod_{k=1}^K \frac{B(h_\phi(k, \cdot) + \beta)}{B(h_\phi(k, -di) + \beta)} \times \prod_{d=1}^D \frac{B(h_\theta(d, \cdot) + \alpha)}{B(h_\theta(d, -di) + \alpha)}$$



$$h_\phi(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]$$

$$h_\theta(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 h_θ(d,2): 5

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 h_φ(1,3): 1
 h_φ(2,3): 2

Collapsed Gibbs Sampling for LDA

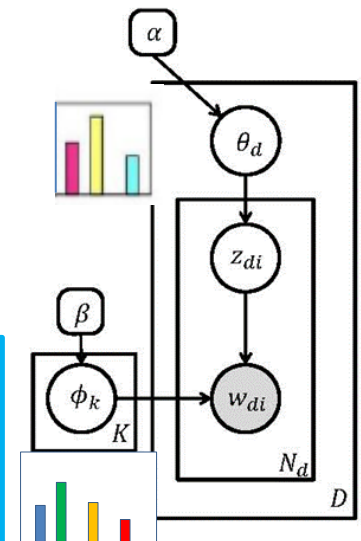
- The Gibbs sampling equation for LDA

$$p(z_{di} | z_{-di}, w, \alpha, \beta) \propto \prod_{k=1}^K \frac{B(h_\phi(k, \cdot) + \beta)}{B(h_\phi(k, -di) + \beta)} \times \prod_{d=1}^D \frac{B(h_\theta(d, \cdot) + \alpha)}{B(h_\theta(d, -di) + \alpha)}$$

$$B(h_\phi(k, \cdot) + \beta) = \frac{\prod_{v=1}^V \Gamma(h_\phi(k, v) + \beta(v))}{\Gamma\left(\sum_{v=1}^V [h_\phi(k, v) + \beta(v)]\right)} \quad \frac{\Gamma(x)}{\Gamma(x-1)} = x - 1$$

$$B(h_\phi(k, -di) + \beta) = \frac{\prod_{v=1}^V \Gamma(h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v))}{\Gamma\left(\sum_{v=1}^V [h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v)]\right)}$$

$$\prod_{k=1}^K \frac{B(h_\phi(k, \cdot) + \beta)}{B(h_\phi(k, -di) + \beta)} \propto \prod_{k=1}^K \prod_{v=1}^V \frac{h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v)}{\sum_{v=1}^V [h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v)]}$$



Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA

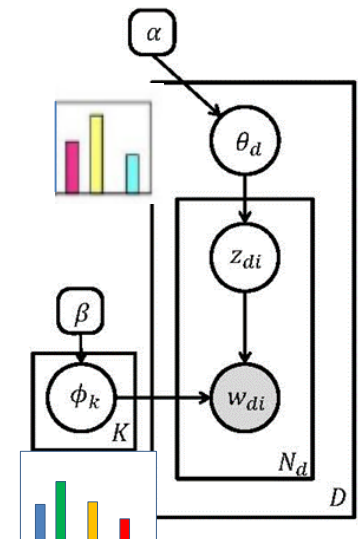
$$p(z_{di} | z_{-di}, w, \alpha, \beta) \propto \prod_{k=1}^K \frac{B(h_\phi(k, \cdot) + \beta)}{B(h_\phi(k, -di) + \beta)} \times \prod_{d=1}^D \frac{B(h_\theta(d, \cdot) + \alpha)}{B(h_\theta(d, -di) + \alpha)}$$

$$\prod_{d=1}^D \frac{B(h_\theta(d, \cdot) + \alpha)}{B(h_\theta(d, -di) + \alpha)} \propto \prod_{d=1}^D \prod_{k=1}^K \frac{h_\theta(d, k) - \delta [z_{di} - k] + \alpha(k)}{\sum_{k=1}^K [h_\theta(d, k) - \delta [z_{di} - k] + \alpha(k)]}$$

$$= \prod_{d=1}^D \prod_{k=1}^K \frac{h_\theta(d, k) - \delta [z_{di} - k] + \alpha(k)}{\sum_{k=1}^K [h_\theta(d, k) + \alpha(k)] - 1}$$

constant

k depends on the sampled z_{di}



Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA

$$p(z_{di} | z_{-di}, w, \alpha, \beta) \propto \prod_{k=1}^K \prod_{v=1}^V = \frac{h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v)}{\sum_{v=1}^V [h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v)]}$$

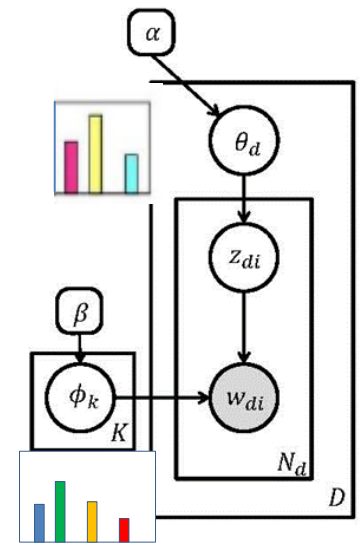
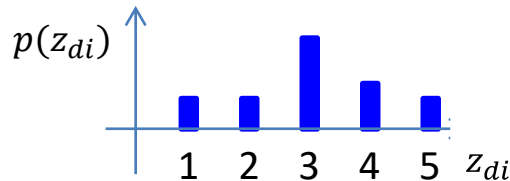
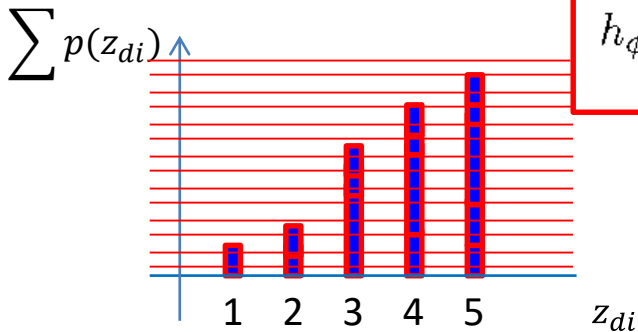
Tractable

$$\times \prod_{d=1}^D \prod_{k=1}^K \{h_\theta(d, k) - \delta[z_{di} - k] + \alpha(k)\},$$

- Sampling

$$h_\theta(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

$$h_\phi(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]$$



What is the inferred label of w_{di} ?

Collapsed Gibbs Sampling for LDA

- Estimation of the multinomial parameters θ and ϕ

Multinomial Dirichlet

$$p(\theta_d | z, \alpha) = \frac{\overbrace{p(z | \theta_d)} \overbrace{p(\theta_d | \alpha)}}{p(z | \alpha)}$$

$$= \text{Dir}(\theta_d | h_\theta(d, \cdot) + \alpha),$$

$$p(\phi_k | z, w, \beta) = \text{Dir}(\phi_k | h_\phi(k, \cdot) + \beta).$$

w: 1, 2, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, 1, 1, 1, 2, 1, 2
 $h_\theta(d, 2): 4$

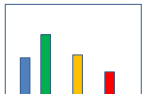
w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, **2**, 1, 1, 2, 1, 2
 $h_\theta(d, 2): 5$

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, **2**, 1, 1, 2, 1, 2
 $h_\phi(1, 3): 1 \leftarrow 2$
 $h_\phi(2, 3): 2 \leftarrow 1$



$$\hat{\theta}_d(k) = E[\theta_d(k) | h_\theta(d, \cdot) + \alpha] = \frac{h_\theta(d, k) + \alpha(k)}{\sum_{k=1}^K [h_\theta(d, k) + \alpha(k)]},$$

$$\hat{\phi}_k(v) = E[\phi_k(v) | h_\phi(k, \cdot) + \beta] = \frac{h_\phi(k, v) + \beta(v)}{\sum_{v=1}^V [h_\phi(k, v) + \beta(v)]}.$$



Interim Summary

- LDA Model (Topic Modelling)
 - Inference of LDA Model
 - Markov Chain Monte Carlo (MCMC)
 - Gibbs Sampling
 - Collapsed Gibbs Sampling for LDA
 - Estimation of Multinomial Parameters via Dirichlet Prior
-

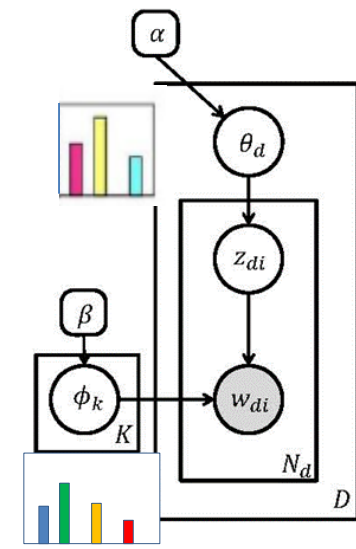
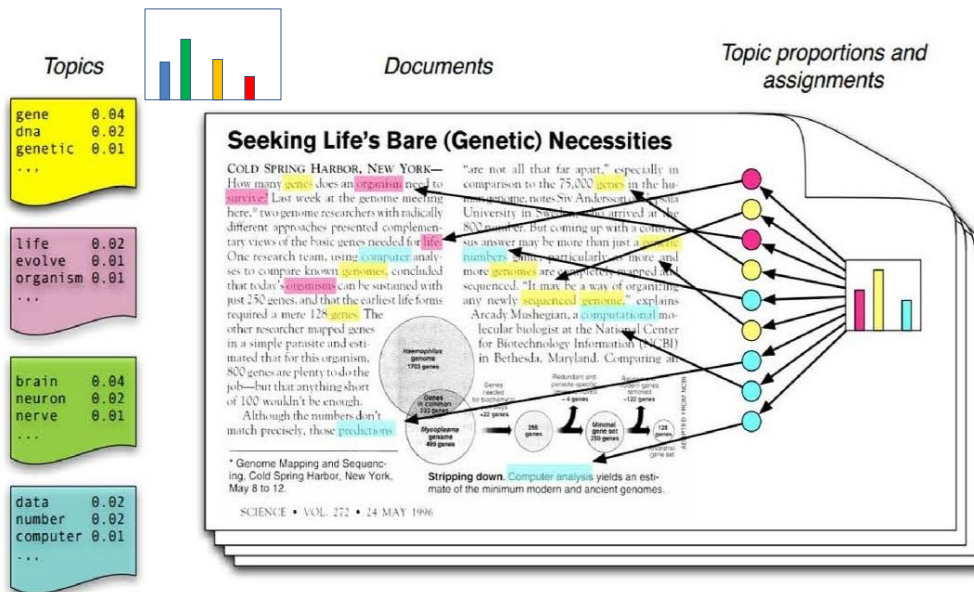
Inference of Bayesian Network: Variational Inference

Jin Young Choi

Outline

- What is variational inference ?
 - Kullback–Leibler divergence (KL-divergence) formulation
 - Dual of KL-divergence
 - Variational Inference for LDA
 - Estimating variational parameters
 - Estimating LDA parameters
 - Application of VI to Generative Image Modeling (**VAE**)
 - Application of LDA to Traffic Pattern Analysis
-

LDA Model (Topic Modelling)



$$p(\phi, \theta, z, w | \alpha, \beta) = \left(\prod_{k=1}^K p(\phi_k | \beta) \right) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{i=1}^{N_d} p(z_{di} | \theta_d) p(w_{di} | z_{di}, \phi)$$

$$\hat{\phi}, \hat{\theta}, \hat{z} = \arg \max_{\phi, \theta, z} p(\phi, \theta, z | w, \alpha, \beta)$$

Likelihood: $p(w | z, \theta, \phi, \alpha, \beta)$

Posteriori: $p(z, \theta, \phi | w, \alpha, \beta)$

Markov Chain Monte Carlo (MCMC)

- Markov Chain Monte Carlo (MCMC) framework

Posteriors

$$p(\theta_d | z, \alpha) = \frac{\overbrace{p(z|\theta_d)}^{\text{Multinomial}} \overbrace{p(\theta_d|\alpha)}^{\text{Dirichlet}}}{p(z|\alpha)}$$

$$= \text{Dir}(\theta_d | h_{\theta}(d, \cdot) + \alpha),$$

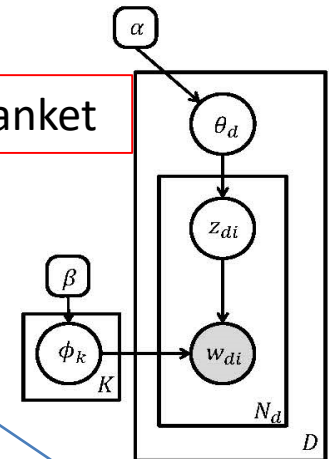
$$p(\phi_k | z, w, \beta) = \text{Dir}(\phi_k | h_{\phi}(k, \cdot) + \beta).$$

$$h_{\theta}(d, k) = \sum_{i=1}^{N_d} \delta [z_{di} - k]$$

$$h_{\phi}(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta [w_{di} - v] \delta [z_{di} - k]$$

Markov Blanket

$$p(z|w, \phi_k, \theta_k)$$

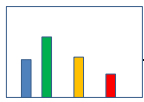


$$\hat{\theta}_d(k) = E[\theta_d(k) | h_{\theta}(d, \cdot) + \alpha] = \frac{h_{\theta}(d, k) + \alpha(k)}{\sum_{k=1}^K [h_{\theta}(d, k) + \alpha(k)]}$$

$$\hat{\phi}_k(v) = E[\phi_k(v) | h_{\phi}(k, \cdot) + \beta] = \frac{h_{\phi}(k, v) + \beta(v)}{\sum_{v=1}^V [h_{\phi}(k, v) + \beta(v)]}$$

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 h_θ(d,2): 5

w: 1, 3, 2, 3, 3, 5, 4, 1, 6
 z: 1, 2, 2, 2, 1, 1, 2, 1, 2
 h_φ(1,3): 1
 h_φ(2,3): 2



Collapsed Gibbs Sampling for LDA

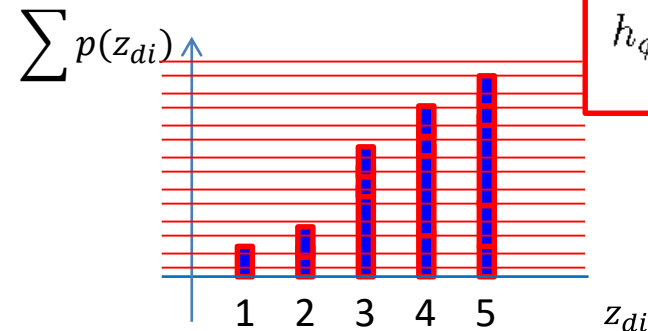
- The Gibbs sampling equation for LDA

$$p(z_{di} | z_{-di}, w, \alpha, \beta) \propto \prod_{k=1}^K \prod_{v=1}^V = \frac{h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v)}{\sum_{v=1}^V [h_\phi(k, v) - \delta[w_{di} - v] \delta[z_{di} - k] + \beta(v)]}$$

Tractable

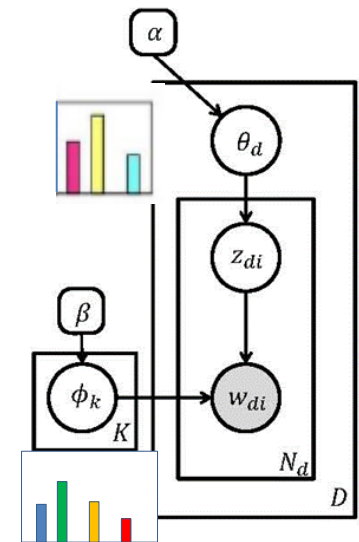
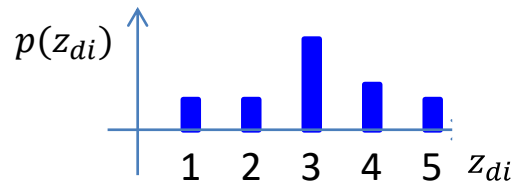
$$\times \prod_{d=1}^D \prod_{k=1}^K \{h_\theta(d, k) - \delta[z_{di} - k] + \alpha(k)\},$$

- Sampling



$$h_\theta(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k]$$

$$h_\phi(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]$$



What is the inferred label of w_{di} ?

Variational Inference (VI)

- Approximating **posteriors** in an **tractable** probabilistic model
- MCMC: **stochastic inference**
- VI: **deterministic inference**
- The posterior distribution $p(z|x)$ can be approximated by a **variational distribution** $q(z)$

$$p(z|x) \approx q(z)$$

where $q(z)$ should belong to a family of **simpler form** than $p(z|x)$

- Minimizing Kullback–Leibler divergence (**KL-divergence**)

$$D[q(z)||p(z|x)] \triangleq \int_z q(z) \log \frac{q(z)}{p(z|x)} dz$$

$$p(\phi|w, z) \propto p(w|\phi, z)p(\phi|\alpha) \approx q(\phi|\gamma)$$
$$C \phi_1^{h(1)+\alpha_1} \phi_2^{h(2)+\alpha_2} \dots \phi_v^{h(v)+\alpha_v} \approx C \phi_1^{\gamma_1} \phi_2^{\gamma_2} \dots \phi_v^{\gamma_v}$$

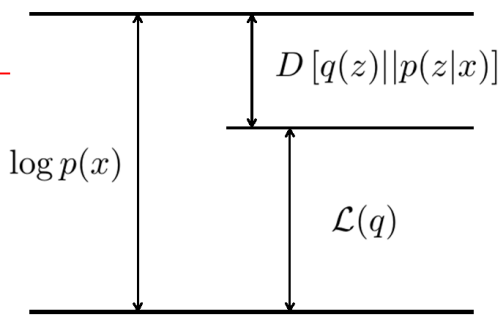
Likelihood: $p(w|z, \theta, \phi, \alpha, \beta)$

Posteriori: $p(z, \theta, \phi|w, \alpha, \beta)$

Variational Inference (VI)

- Kullback–Leibler divergence (KL-divergence)

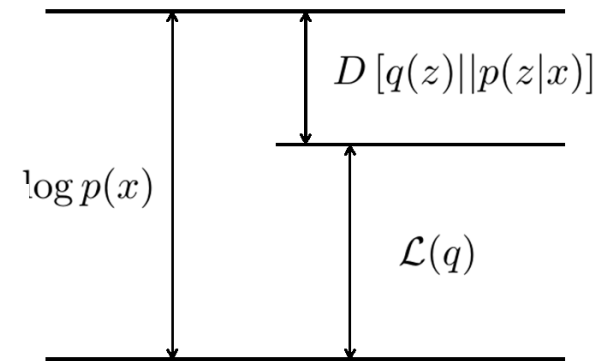
$$\begin{aligned}
 D[q(z)||p(z|x)] &= \int_z q(z) \log \frac{q(z)p(x)}{p(z,x)} dz && \boxed{p(z|x) = \frac{p(z,x)}{p(x)}} \\
 &= \int_z q(z) \log \frac{q(z)}{p(z,x)} dz + \int_z q(z) \log p(x) dz \\
 &= \int_z q(z) \log \frac{q(z)}{p(z,x)} dz + \log p(x) \underbrace{\int_z q(z) dz}_{=1}
 \end{aligned}$$

$$\begin{aligned}
 &= \int_z q(z) \log \frac{q(z)}{p(z,x)} dz + \log p(x). \\
 \text{Minimize } &\boxed{-\mathcal{L}(q) \triangleq} \quad \boxed{\text{constant}}
 \end{aligned}$$


Variational Inference (VI)

- Dual of KL-divergence

$$\begin{aligned}\mathcal{L}(q) &\triangleq \int_z q(z) \log \frac{p(z, x)}{q(z)} dz \\ &= \int_z q(z) \log p(z, x) dz - \int_z q(z) \log q(z) dz \\ &= E_q [\log p(z, x)] + \underbrace{H [q(z)]}_{\text{entropy of } q(z)}.\end{aligned}$$



Variational Inference (VI)

- Choose a variational distribution $q(z)$
- An approximation (Parisi, 1988) was proposed

$$q(z) = \prod_{i=1}^N q_i(z_i) = \prod_{i=1}^N q(z_i | \lambda_i)$$

where λ_i is a variational parameter for each hidden variable z_i

- Estimation of λ $\hat{\lambda} = \max_{\lambda} \bar{\mathcal{L}}(\lambda)$

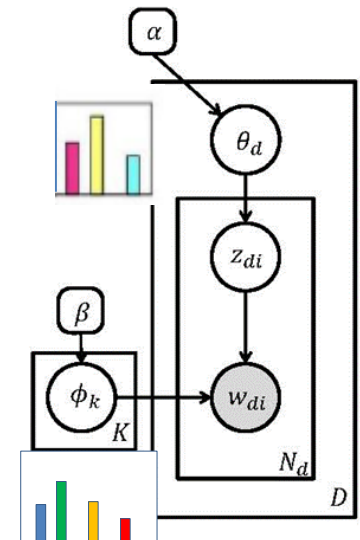
$$\nabla_{\lambda} \bar{\mathcal{L}}(\lambda) = 0$$

where $q(\cdot)$ would be designed for $\bar{\mathcal{L}}(\lambda)$ to be convex.

$z \rightarrow \phi$	$p(\phi w, z) \propto p(w \phi, z)p(\phi \alpha) \approx q(\phi \gamma)$
$x \rightarrow w$	$C\phi_1^{h(1)+\alpha_1}\phi_2^{h(2)+\alpha_2}\dots\phi_V^{h(V)+\alpha_v} \approx C\phi_1^{\gamma_1}\phi_2^{\gamma_2}\dots\phi_V^{\gamma_V}$
$z \rightarrow \theta$	$p(\theta z) \propto p(z \theta)p(\theta \alpha) \approx q(\theta \mu)$
$x \rightarrow z$	$C\theta_1^{h(1)+\alpha_1}\theta_2^{h(2)+\alpha_2}\dots\theta_K^{h(K)+\alpha_v} \approx C\theta_1^{\mu_1}\theta_2^{\mu_2}\dots\theta_K^{\mu_K}$

$\bar{\mathcal{L}}(\gamma)$

$\bar{\mathcal{L}}(\mu)$



Variational Inference for LDA

- Objective function and joint probability of LDA

$$\phi^*, \theta^*, z^* = \arg \max_{\phi, \theta, z} p(\phi, \theta, z | w, \alpha, \beta)$$

$$= \arg \max_{\phi, \theta, z} \frac{p(\phi, \theta, z, w | \alpha, \beta)}{p(w | \alpha, \beta)}$$

$$= \arg \max_{\phi, \theta, z} p(\phi, \theta, z, w | \alpha, \beta),$$

$$\approx q(\phi, \theta, z | \lambda, \varphi, \gamma)$$

Maximization is intractable
→ Variational Inference

where

$$p(\phi, \theta, z, w | \alpha, \beta) = \left(\prod_{k=1}^K p(\phi_k | \beta) \right) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{i=1}^{N_d} p(z_{di} | \theta_d) p(w_{di} | z_{di}, \phi)$$

$$q(\phi, \theta, z | \lambda, \varphi, \gamma) =$$

$$\left(\prod_{k=1}^K q(\phi_k | \lambda_k) \right) \left(\prod_{d=1}^D q(\theta_d | \gamma_d) \right) \left(\prod_{d,i}^{D, N_d} q(z_{di} | \varphi_{di}) \right)$$

Variational Inference for LDA

- A simpler variational distribution

$$q(\phi, \theta, z | \lambda, \varphi, \gamma) =$$

$$\left(\prod_{k=1}^K q(\phi_k | \lambda_k) \right) \left(\prod_{d=1}^D q(\theta_d | \gamma_d) \right) \left(\prod_{d,i}^{D, N_d} q(z_{di} | \varphi_{di}) \right)$$

where λ, γ, φ are the variational parameters used for approximate inference of ϕ, θ, z , respectively.

$$\phi_k | \lambda_k \sim \text{Dirichlet}(\phi_k | \lambda_k) \quad \text{Dir}(\phi_k | h_\phi(k, \cdot) + \beta).$$

$$\theta_d | \gamma_d \sim \text{Dirichlet}(\theta_d | \gamma_d) \quad \longrightarrow \quad \text{Dir}(\theta_d | h_\theta(d, \cdot) + \alpha)$$

$$z_{di} | \varphi_{di} \sim \text{Multi}(z_{di} | \varphi_{di}). \quad \text{Multi}(z_{di} | \theta_{\cdot d}, w_{di})$$

$$p(\theta | z) \propto p(z | \theta) p(\theta | \alpha) \approx q(\theta | \gamma)$$

$$C \theta_1^{h(1)+\alpha_1} \theta_2^{h(2)+\alpha_2} \dots \theta_K^{h(K)+\alpha_K} \approx C \theta_1^{\gamma_1} \theta_2^{\gamma_2} \dots \theta_K^{\gamma_K}$$

$$p(\phi | w, z) \propto p(w | \phi, z) p(\phi | \beta) \approx q(\phi | \lambda)$$

$$C \phi_1^{h(1)+\beta_1} \phi_2^{h(2)+\beta_2} \dots \phi_V^{h(V)+\beta_V} \approx C \phi_1^{\lambda_1} \phi_2^{\lambda_2} \dots \phi_V^{\lambda_V}$$

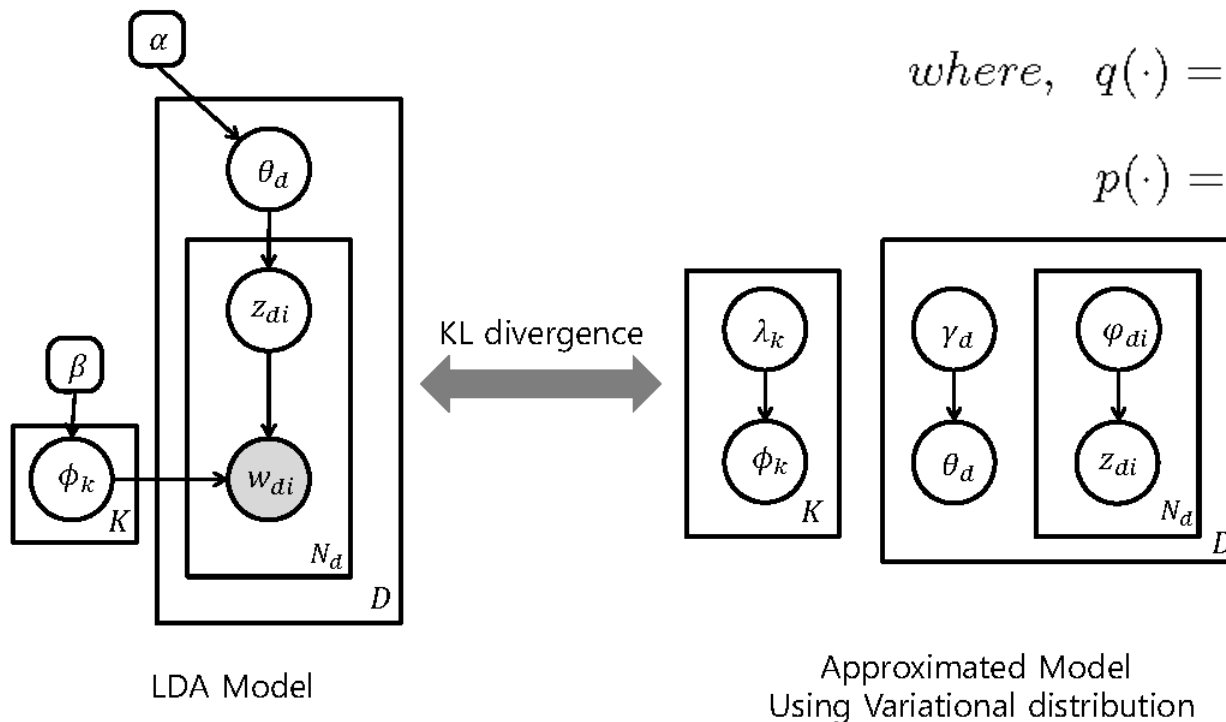
Variational Inference for LDA

- Optimal values of the variational parameters

$$\lambda^*, \gamma^*, \varphi^* = \arg \min_{\lambda, \gamma, \varphi} D [q(\cdot) \| p(\cdot)],$$

$$\text{where, } q(\cdot) = q(\phi, \theta, z | \lambda, \gamma, \varphi),$$

$$p(\cdot) = p(\phi, \theta, z | w, \alpha, \beta).$$



$$p(\theta | z) \propto p(z | \theta) p(\theta | \alpha) \approx q(\theta | \gamma)$$

$$C \theta_1^{h(1)+\alpha_1} \theta_2^{h(2)+\alpha_2} \dots \theta_K^{h(K)+\alpha_K} \approx C \theta_1^{\gamma_1} \theta_2^{\gamma_2} \dots \theta_K^{\gamma_K}$$

$$p(\phi | w, z) \propto p(w | \phi, z) p(\phi | \beta) \approx q(\phi | \lambda)$$

$$C \phi_1^{h(1)+\beta_1} \phi_2^{h(2)+\beta_2} \dots \phi_V^{h(V)+\beta_V} \approx C \phi_1^{\lambda_1} \phi_2^{\lambda_2} \dots \phi_V^{\lambda_V}$$

Variational Inference for LDA

- Optimal values of the variational parameters

$$\lambda^*, \gamma^*, \varphi^* = \arg \min_{\lambda, \gamma, \varphi} D [q(\cdot) \| p(\cdot)],$$

$$\phi^*, \theta^*, z^* = \arg \max_{\phi, \theta, z} p(\phi, \theta, z | w, \alpha, \beta)$$

$$\text{where, } q(\cdot) = q(\phi, \theta, z | \lambda, \gamma, \varphi),$$

$$p(\cdot) = p(\phi, \theta, z | w, \alpha, \beta).$$

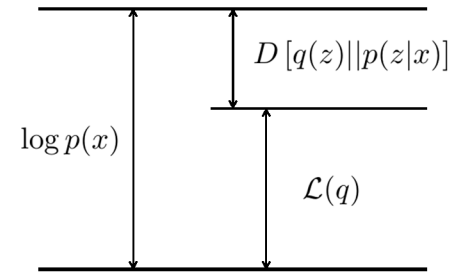
$$\text{where } p(\phi, \theta, z | w, \alpha, \beta) = p(\phi, \theta, z, w | \alpha, \beta) / p(w)$$

$$p(\phi, \theta, z, w | \alpha, \beta) = \left(\prod_{k=1}^K p(\phi_k | \beta) \right) \prod_{d=1}^D p(\theta_d | \alpha) \prod_{i=1}^{N_d} p(z_{di} | \theta_d) p(w_{di} | z_{di}, \phi)$$

$$q(\phi, \theta, z | \lambda, \varphi, \gamma) =$$

$$\left(\prod_{k=1}^K q(\phi_k | \lambda_k) \right) \left(\prod_{d=1}^D q(\theta_d | \gamma_d) \right) \left(\prod_{d,i}^{D, N_d} q(z_{di} | \varphi_{di}) \right)$$

Variational Inference for LDA



- Dual of KL divergence

$$\mathcal{L}(q) = \log p(w|\alpha, \beta) - D[q(\phi, \theta, z|\lambda, \gamma, \varphi) || p(\phi, \theta, z|w, \alpha, \beta)]$$

- Yielding the optimal value of each variational parameter

$$\frac{\partial \mathcal{L}(q)}{\partial \varphi_{di}} = 0 \quad \varphi_{di}(k) \propto \exp \left\{ E_q[\log \theta_d(k)] + E_q[\log \phi_k(w_{di})] \right\}$$

$$\frac{\partial \mathcal{L}(q)}{\partial \gamma_d} = 0 \quad \gamma_d(k) = \alpha(k) + \sum_{i=1}^{N_d} \varphi_{di}(k) \quad \begin{cases} h_\theta(d, k) = \sum_{i=1}^{N_d} \delta[z_{di} - k] \\ h_\phi(k, v) = \sum_{d=1}^D \sum_{i=1}^{N_d} \delta[w_{di} - v] \delta[z_{di} - k]. \end{cases}$$

$$\frac{\partial \mathcal{L}(q)}{\partial \lambda_k} = 0. \quad \lambda_k(v) = \beta(v) + \sum_{d=1}^D \sum_{i=1}^{N_d} \varphi_{di}(k) \delta[w_{di} - v].$$

$$\phi_k | \lambda_k \sim \text{Dirichlet}(\phi_k | \lambda_k)$$

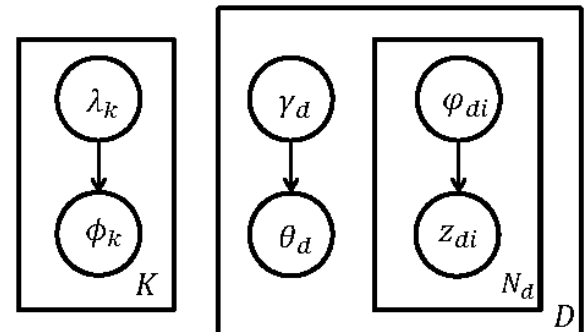
$$\theta_d | \gamma_d \sim \text{Dirichlet}(\theta_d | \gamma_d)$$

$$z_{di} | \varphi_{di} \sim \text{Multi}(z_{di} | \varphi_{di}).$$

$$\text{Dir}(\phi_k | h_\phi(k, \cdot) + \beta).$$

$$\text{Dir}(\theta_d | h_\theta(d, \cdot) + \alpha)$$

$$\text{Multi}(z_{di} | \theta_{d, \cdot}, w_{di})$$

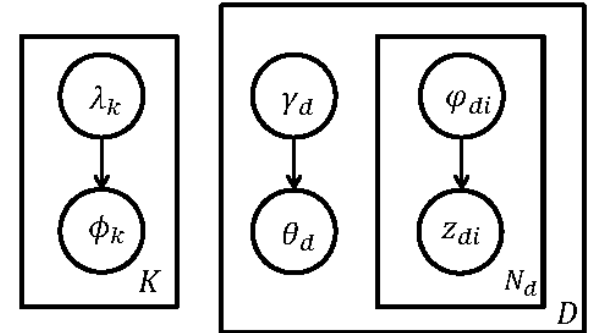


Variational Inference for LDA

- Expectation

$$E_q [\log \theta_d(k)] = \Psi(\gamma_d(k)) - \Psi\left(\sum_{k=1}^K \gamma_d(k)\right)$$

$$E_q [\log \phi_k(w_{di})] = \Psi(\lambda_k(w_{di})) - \Psi\left(\sum_{v=1}^V \lambda_k(v)\right)$$



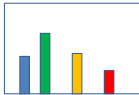
where $\Psi(\cdot)$ is the digamma function (Blei et al. 2003) given by

$$\Psi(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$


- The **iterative updates** of the variational parameters are guaranteed to **converge into a stationary point**
- For the iteration, **ϕ** are updated with **λ and γ fixed**, and **λ and γ** are updated given the fixed **ϕ**

Variational Inference for LDA

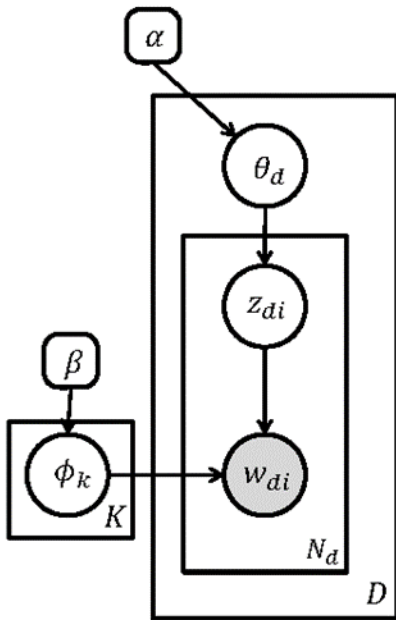
- The final distribution results ϕ and θ



$$\hat{\phi}_k(v) = E_{\underbrace{q(\phi_k|\lambda_k)}_{\text{Dirichlet}}} [\phi_k(v)|\lambda_k] = \frac{\lambda_k(v)}{\sum_{v=1}^V \lambda_k(v)}$$

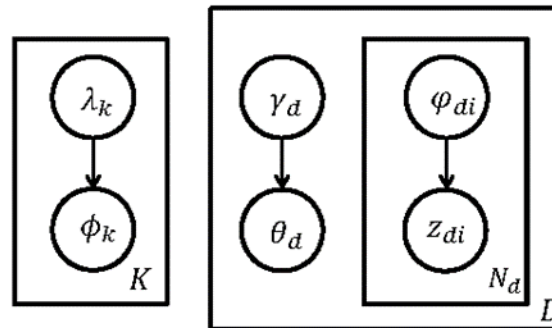


$$\hat{\theta}_d(k) = E_{\underbrace{q(\theta_d|\gamma_d)}_{\text{Dirichlet}}} [\theta_d(k)|\gamma_d] = \frac{\gamma_d(k)}{\sum_{k=1}^K \gamma_d(k)}$$

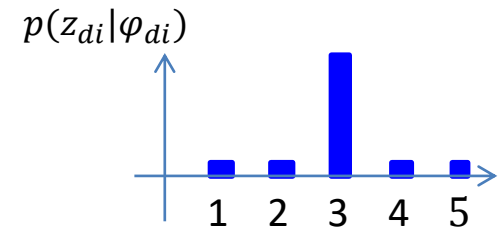


LDA Model

KL divergence



Approximated Model
Using Variational distribution



$$\phi_k|\lambda_k \sim \text{Dirichlet}(\phi_k|\lambda_k) \quad \text{Dir}(\phi_k|h_\phi(k, \cdot) + \beta).$$

$$\theta_d|\gamma_d \sim \text{Dirichlet}(\theta_d|\gamma_d) \quad \text{Dir}(\theta_d|h_\theta(d, \cdot) + \alpha)$$

$$z_{di}|\phi_{di} \sim \text{Multi}(z_{di}|\phi_{di}). \quad \text{Multi}(z_{di}|\theta_d, w_{di})$$

References for LDA and VI

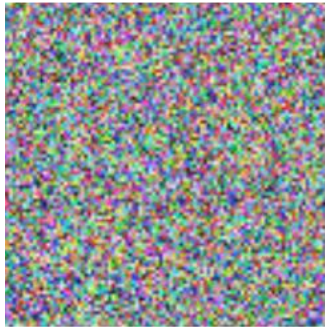
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *JML. Res.*, 3, 993–1022.
- Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4), 77–84.
- Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1(1), 203–232.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2), 183–233.

URL <http://dx.doi.org/10.1023/A:1007665907178>

Generative Image Modeling

Ex) Image Generation

Noise $\sim N(0,1)$



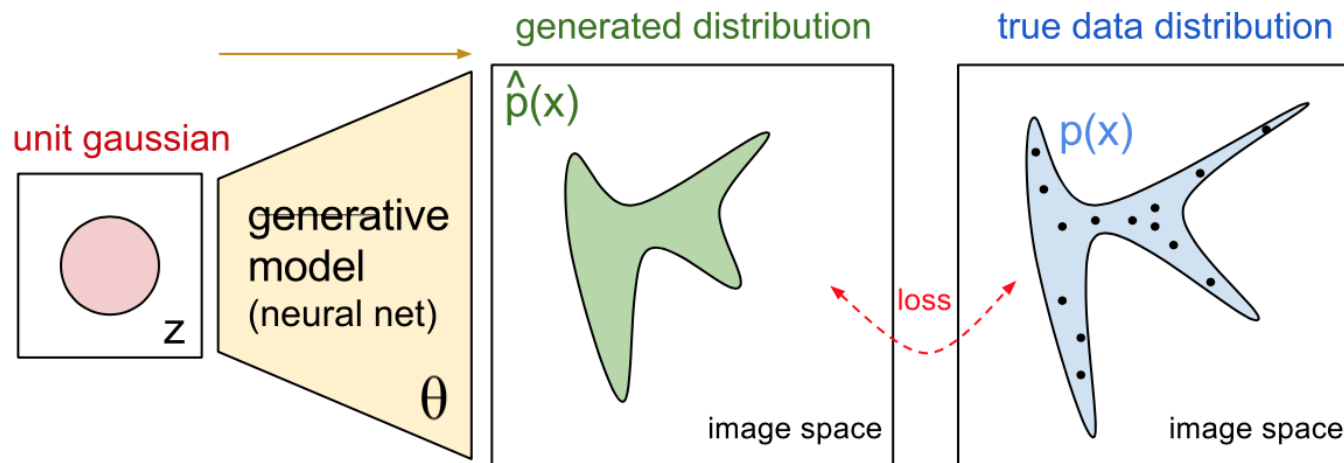
Generative
Model



Generative Image Modeling

Goal: Model the distribution $p(x)$

- cf) Discriminative approach

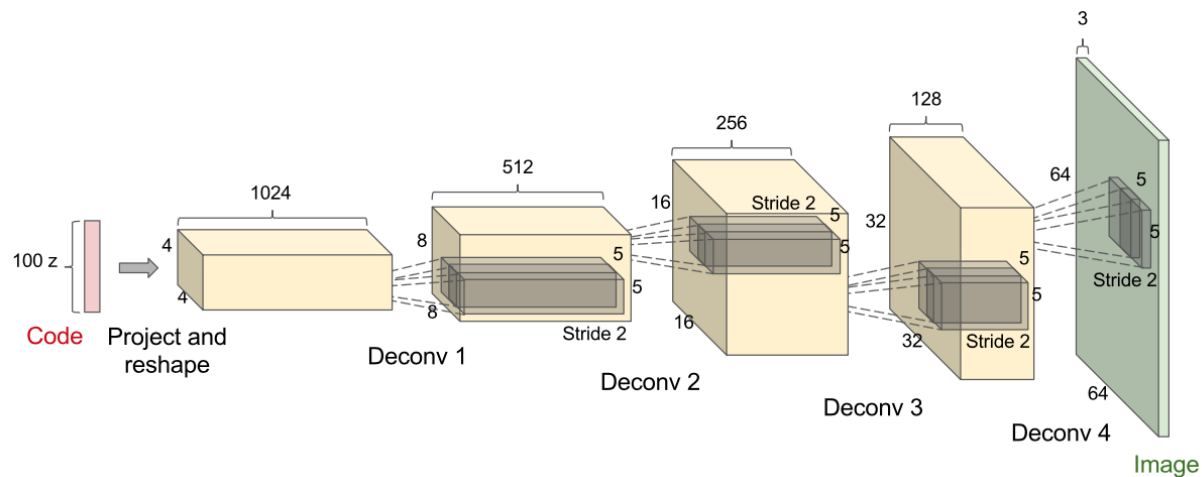


Issue: High Dimensional
low resolution ($64 \times 64 \times 3 = 12,288$)

Generative Image Modeling

Goal: Model the distribution $p(x)$

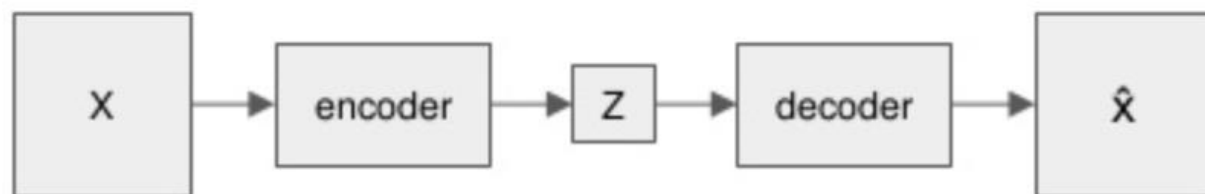
- Using Deep Neural Networks (CNNs)



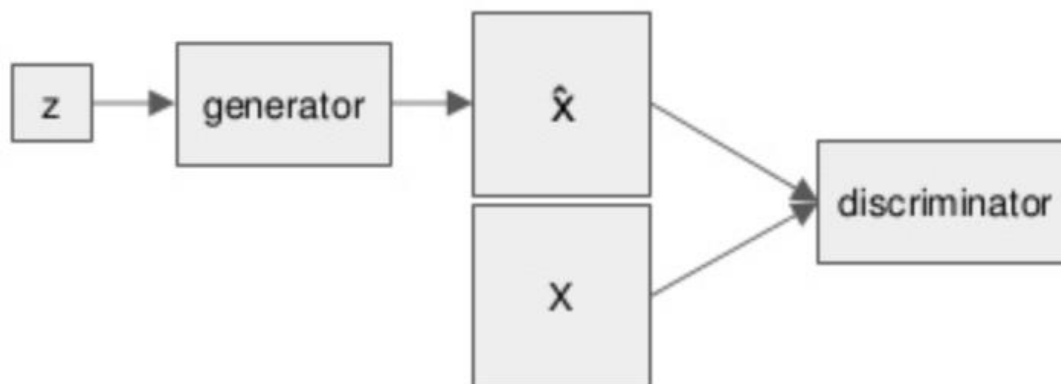
Deep Generative Models

Two prominent approaches

- Variational Auto-encoder (VAE)
- Generative Adversarial Networks (GAN)

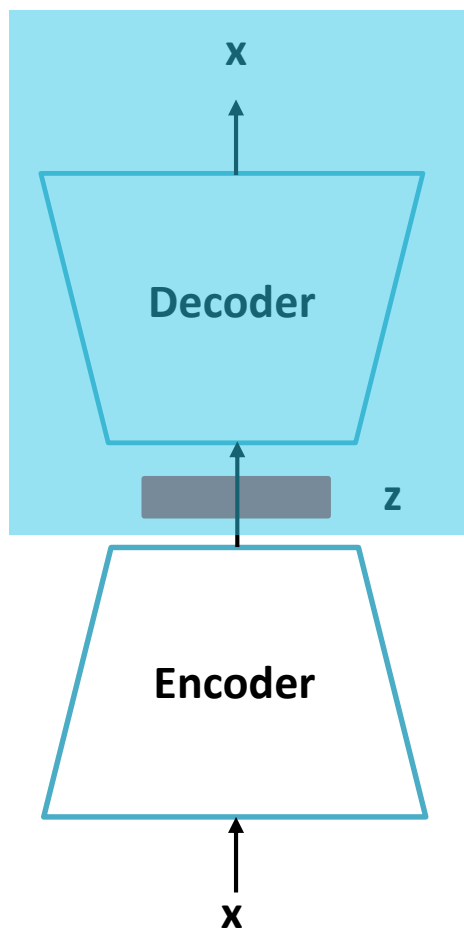


Variational
Autoencoders (VAE)
[Kingma and Welling
\[1312.6114\]](#)



Generative Adversarial
Networks (GAN)
[Goodfellow et al. \[1406.2661\]](#)

Variational Auto-encoder (VAE)



Reconstruction Loss

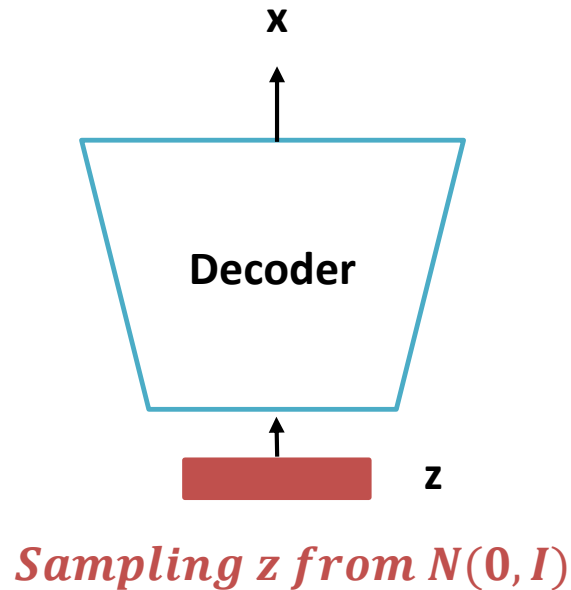
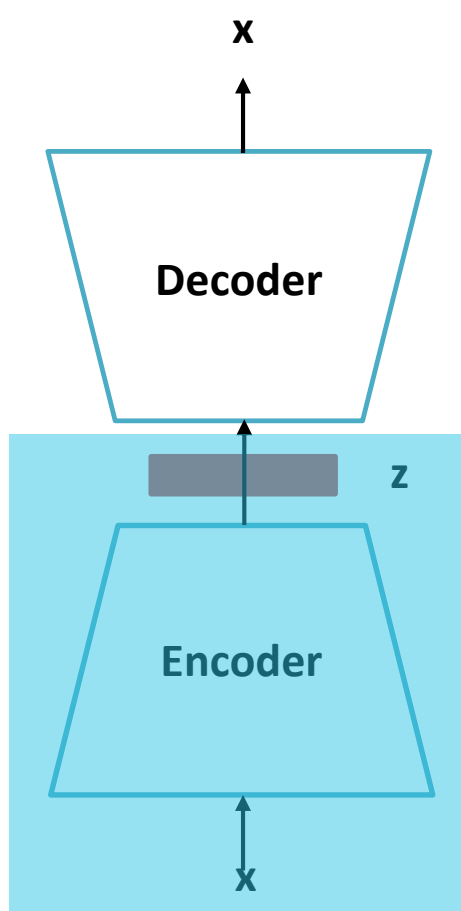
Variational Inference

$$Loss = -\log P_{\theta}(x|z) + D_{KL}(q_{\phi}(z|x) || P_{\theta}(z))$$

$p_{\theta}(x|z)$: a multivariate Gaussian (real-valued data)
a Bernoulli (binary-valued data)



Variational Auto-encoder (VAE)



Variational Inference

$$\text{Loss} = -\log P_{\theta}(x|z) + D_{KL}(q_{\phi}(z|x) || P_{\theta}(z))$$

$$p_{\theta}(z) \sim N(0, I)$$

Interim Summary

- What is variational inference ?
 - Kullback–Leibler divergence (KL-divergence) formulation
 - Dual of KL-divergence
 - Variational Inference for LDA
 - Estimating variational parameters
 - Estimating LDA parameters
 - Application of VI to Generative Image Modeling
-