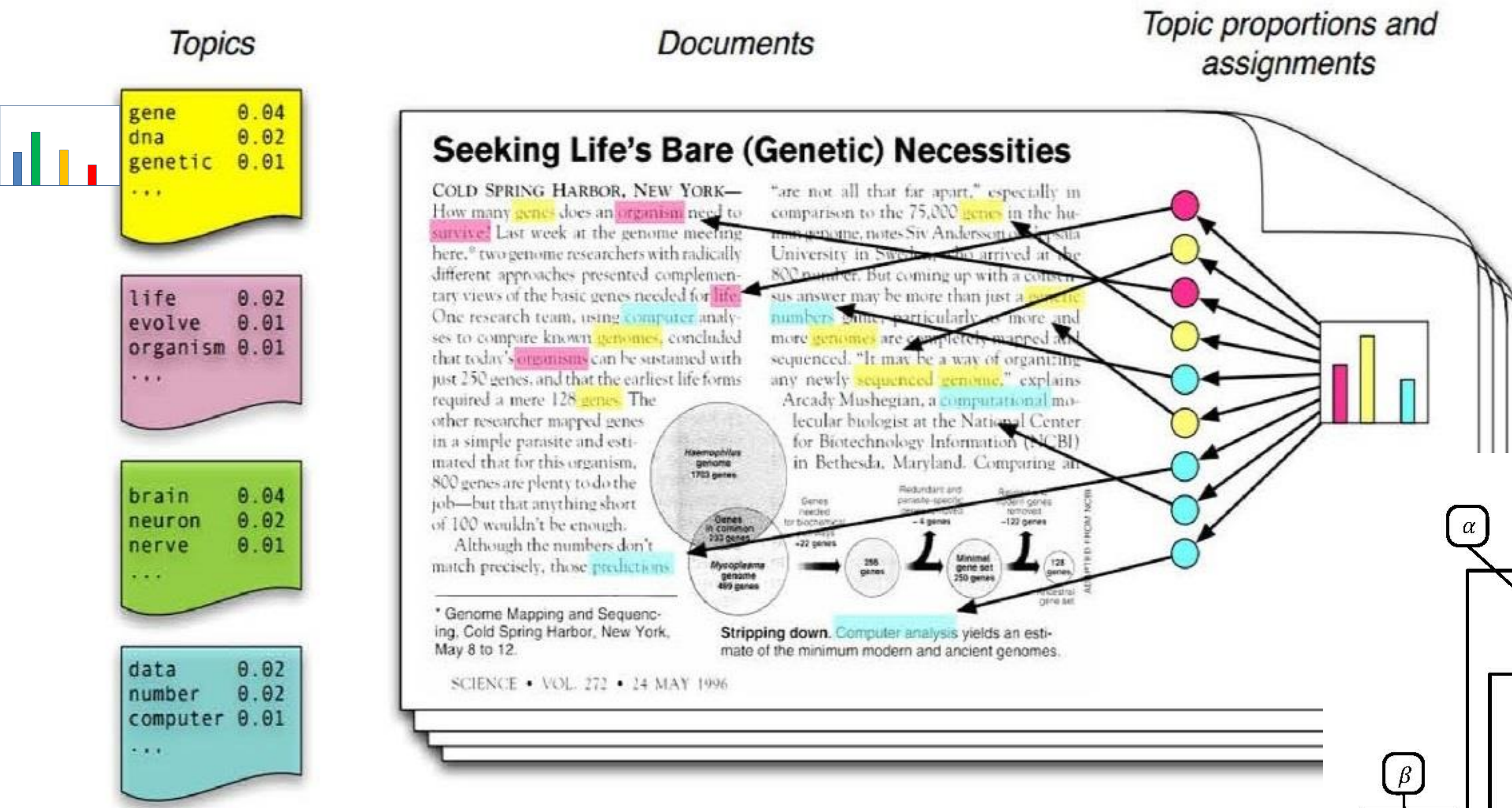
Inference of Bayesian Network: MCMC

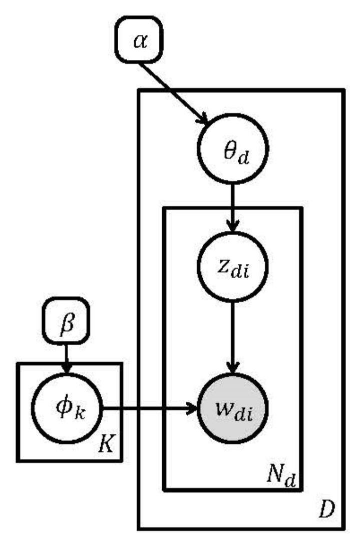
Jin Young Choi

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)



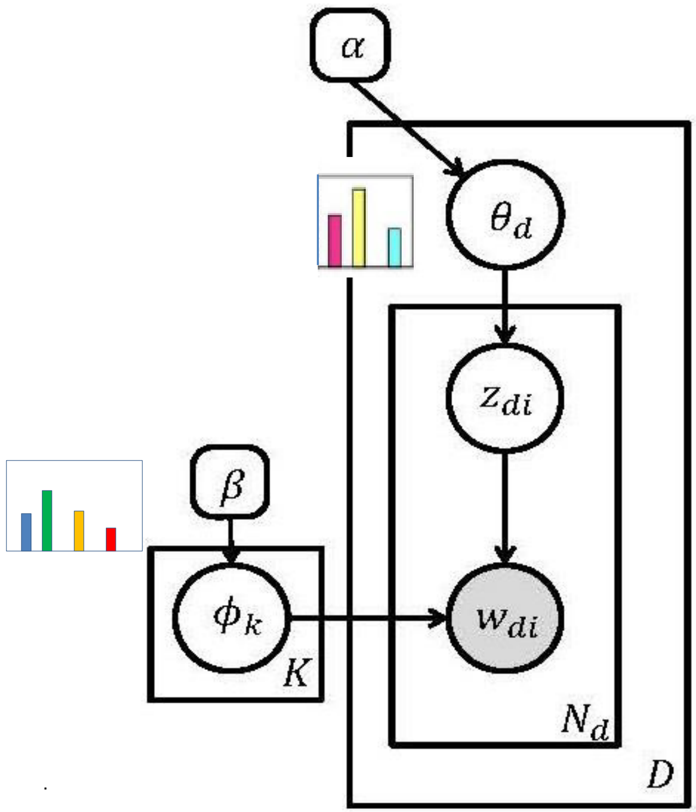


LDA Model

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

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

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



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.

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

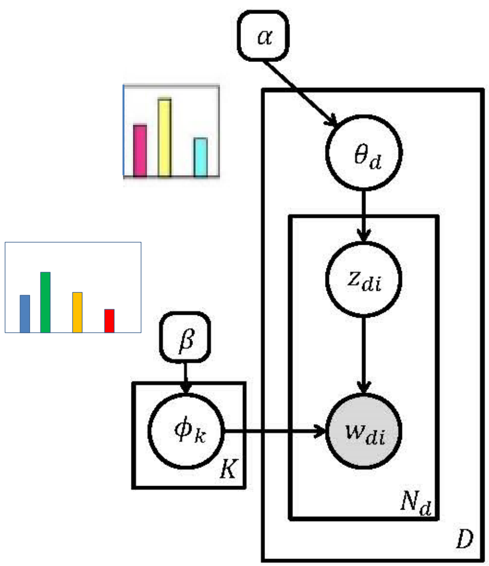
Mathematical description

Choose $\theta_d \sim Dir(\alpha)$.
Choose $\phi_k \sim Dir(\beta)$.
Choose a topic $z_{ji}|\theta_d \sim Multi(\theta_d)$.
Choose a word $w_{ji}|\phi_k, z_{di} \sim Multi(\phi_{z_{di}})$.

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

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



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

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

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

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

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

Inference of LDA Model

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

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

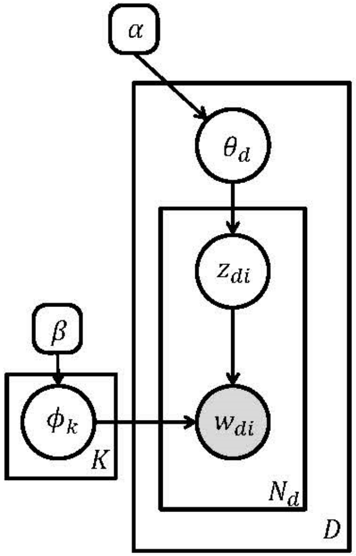
- Maximum A posteriori Probability (MAP) given observation w, α, β

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

Not Convex
Closed-form solution is not available

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

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

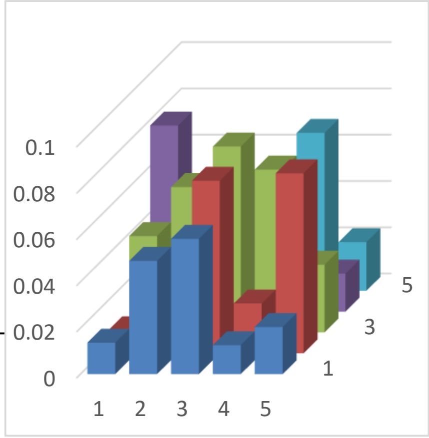


Likelihoods

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

Discrete Variables, Dirichlet

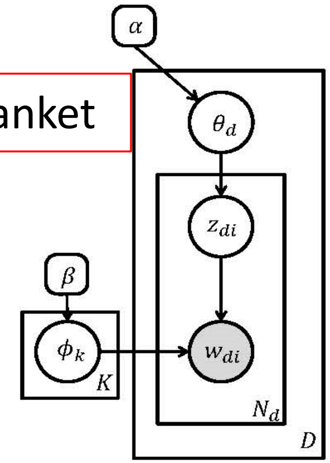
- Parameters: $\alpha_1, \ldots, \alpha_K > 0$ (concentration hyper-parameter)
- Support: $\mu_1, \ldots, \mu_K \in (0,1)$ where $\sum_{i=1}^{K} \mu_i = 1$
- Dir(K, α): $p(\mu_1, \ldots, \mu_K | \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K}(\alpha_i))}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \mu_1^{\alpha_1 - 1} \quad \ldots \mu_K^{\alpha_K - 1}$
- Mul(K, μ): $p(x_1, \ldots, x_K | \mu_1, \ldots, \mu_K) = \frac{n!}{x_1! \ldots x_K!} \mu_1^{x_1} \quad \ldots \mu_K^{x_K}$
- Dir(K, c + α): $p(\mu | x, \alpha) \propto p(x|\mu) p(\mu|\alpha)$
  where $c = (c_1, \ldots, c_K)$ 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

Markov Blanket



Posteriors

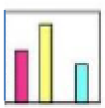
$$p(\theta_d | z, \alpha) = \frac{\overbrace{p(z|\theta_d)}^{Multinomial} \; \overbrace{p(\theta_d|\alpha)}^{Dirichlet}}{p(z|\alpha)}$$

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

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

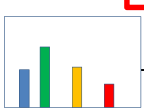
$$h_\theta(d,k) = \sum_{i=1}^{N_d} \delta\left[z_{di} - k\right])$$

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



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

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



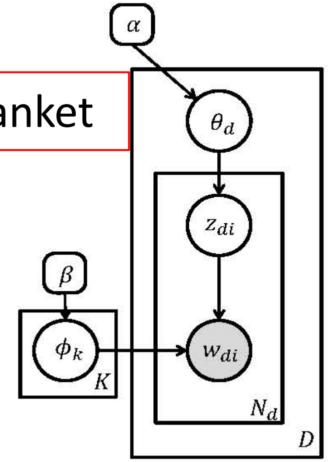
i: 1, 2, 3, 4, 5, 6, 7, 8, 9
w: 1, 3, 2, 3, 3, 5, 4, 1, 6
z : 1, 2, 2, 2, 1, 1, 2, 1, 2
h_θ(d,2): 5

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_ϕ(1,3): 1
h_ϕ(2,3): 2

Markov Chain Monte Carlo (MCMC)

- Markov Chain Monte Carlo (MCMC) framework

Markov Blanket



Posteriors

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

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

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

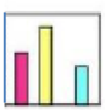
$$h_\theta(d,k) = \sum_{i=1}^{N_d} \delta[z_{di} - k])$$

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

Gibbs sampling

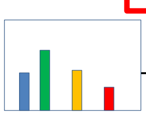
$p(z|w,\phi_k,\alpha)$

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



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

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



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

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

Gibbs Sampling (Review)

$p(z) = p(z_1, z_2, ..., z_N)$

1. Randomly initialize each $z_i^1 \in \{1, 2, ..., K\}$, where $i = 1, 2, ..., N$,

2. For each step $t = 1, 2, ..., T$:

- Replace z_1^t by a new value z_1^{t+1}, sampling $z_1^{t+1} \sim p(z_1 | z_2^t, z_3^t, ..., z_N^t)$.
- Replace z_2^t by a new value z_2^{t+1}, sampling $z_2^{t+1} \sim p(z_2 | z_1^{t+1}, z_3^t, ..., z_N^t)$.
- ...
- Replace z_j^t by a new value z_j^{t+1},

  sampling $z_j^{t+1} \sim p(z_j | z_1^{t+1}, ..., z_{j-1}^{t+1}, z_{j+1}^t, ...z_N^t)$.
- Replace z_N^t by a new value z_N^{t+1}, sampling $z_N^{t+1} \sim p(z_N | z_1^{t+1}, ..., z_{N-1}^{t+1})$.

Collapsed Gibbs Sampling for LDA

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

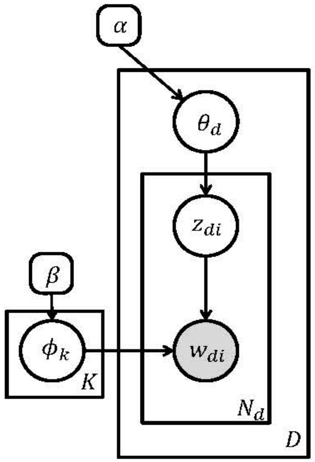
$$p(\theta_d | z, \alpha) = \frac{\overbrace{p(z|\theta_d)}^{Multinomial} \overbrace{p(\theta_d|\alpha)}^{Dirichlet}}{p(z|\alpha)} = Dir(\theta_d | h_\theta(d, \cdot) + \alpha),$$

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

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

Collapsed Gibbs Sampling

- Collapsed Gibbs Sampling
  - θ and $\emptyset$ are induced by the association between z and w
  - z is sufficient statistic to estimate θ and $\emptyset$
  - Simpler algorithm can be used by sampling only z after marginalizing θ and $\emptyset$.
  - $p(z|w, \alpha, \beta) = \iint p(\theta, \emptyset, z | w, \alpha, \beta)\, d_\theta d_\phi$



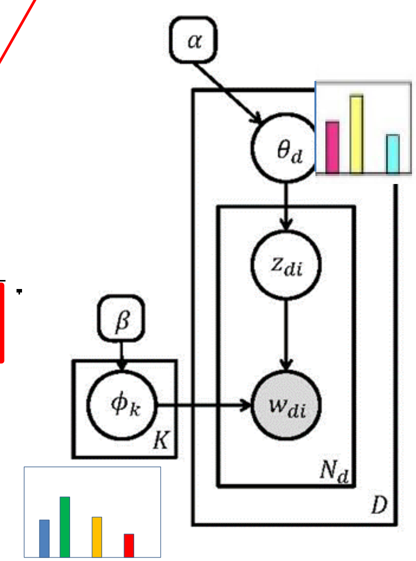
Collapsed Gibbs Sampling for LDA

- 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 $p(w|z,\phi)$

$$p(w|z,\beta) = \int p(\phi|\beta)p(w|z,\phi)d\phi$$

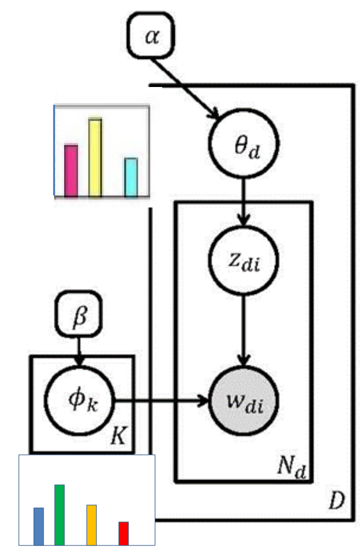
$$= \int \left\{ \prod_{k=1}^{K} p(\phi_k|\beta) \right\} \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(w_{di}|z_{di},\phi)d\phi$$

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

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

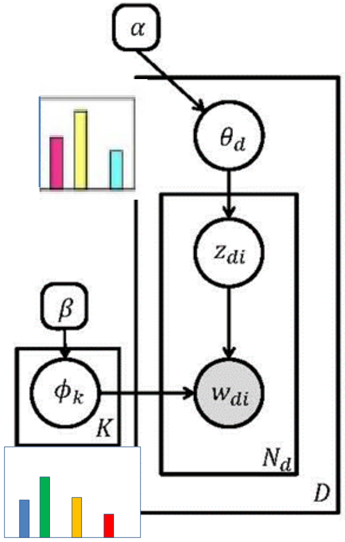
- The number of times that the word $w_{di} = v$ is assigned to the topic $z_{di} = k$

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

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

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

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

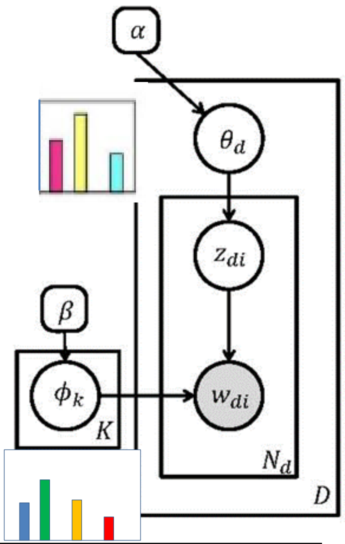
- In consequence

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

$$= \int \left\{ \prod_{k=1}^{K} \frac{1}{B(\beta)} \prod_{v=1}^{V} \phi_k(v)^{\beta(v)-1} \right\} \prod_{k=1}^{K} \prod_{v=1}^{V} \{\phi_k(v)\}^{h_\phi(k,v)} d\phi$$

$$= \prod_{k=1}^{K} \frac{1}{B(\beta)} \int \prod_{v=1}^{V} \{\phi_k(v)\}^{h_\phi(k,v)+\beta(v)-1} d\phi.$$

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



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

Collapsed Gibbs Sampling for LDA

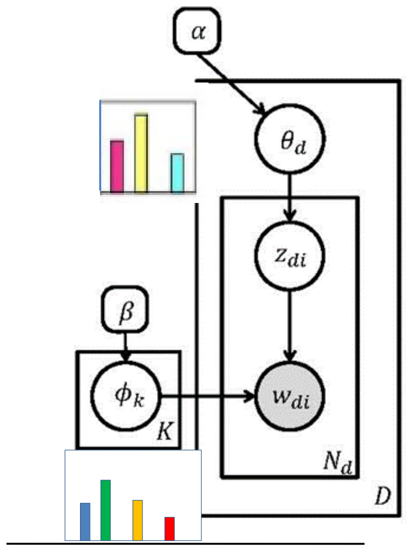
- Integrating PDF

$$p(w|z,\beta) = \prod_{k=1}^{K} \frac{1}{B(\beta)} \int \prod_{v=1}^{V} \{\phi_k(v)\}^{h_\phi(k,v)+\beta(v)-1} d\phi$$

$$= \prod_{k=1}^{K} \frac{B(h_\phi(k,\cdot)+\beta)}{B(\beta)} \underbrace{\int \frac{1}{B(h_\phi(k,\cdot)+\beta)} \prod_{v=1}^{V} \{\phi_k(v)\}^{h_\phi(k,v)+\beta(v)-1} d\phi}_{=1\ (Integral\ of\ pdf)}$$

$$= \prod_{k=1}^{K} \frac{B(h_\phi(k,\cdot)+\beta)}{B(\beta)}.$$

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



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

Topic portion corresponding to a word

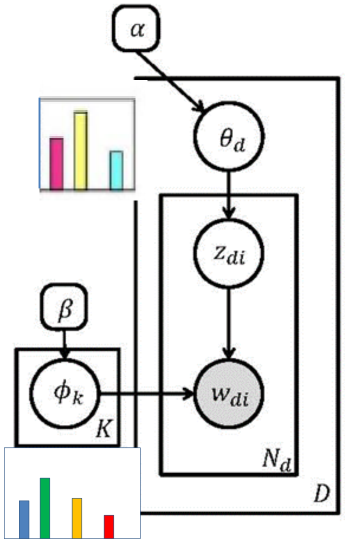
$$p(z|\alpha) = \int p(\theta \mid \alpha) p(z|\theta) d\theta$$

$$= \int \prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{i=1}^{N_d} p(z_{di}|\theta_d) d\theta$$

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

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

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



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

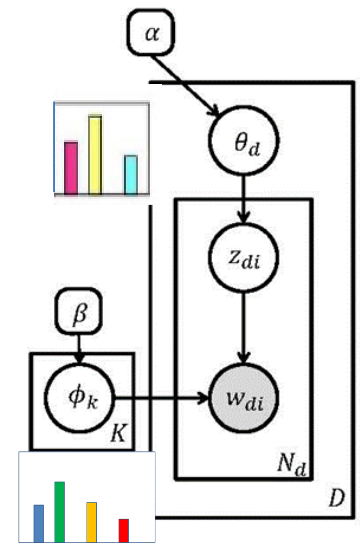
$$p(z|\alpha) = \int \prod_{d=1}^{D} \left\{ \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_d(k)^{\alpha(k)-1} \right\} \prod_{k=1}^{K} \theta_d(k)^{h_\theta(d,k)} d\theta$$

$$= \int \prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_d(k)^{\alpha(k)-1} \theta_d(k)^{h_\theta(d,k)} d\theta$$

$$= \int \prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_d(k)^{h_\theta(d,k)+\alpha(k)-1} d\theta$$

$$= \prod_{d=1}^{D} \frac{B(h_\theta(d,\cdot)+\alpha)}{B(\alpha)} \underbrace{\int \frac{1}{B(h_\theta(d,\cdot)+\alpha)} \prod_{k=1}^{K} \theta_d(k)^{h_\theta(d,k)+\alpha(k)-1} d\theta}_{=1\ (Integral\ of\ pdf)}$$

$$= \prod_{d=1}^{D} \frac{B(h_\theta(d,\cdot)+\alpha)}{B(\alpha)}.$$



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

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

Collapsed Gibbs Sampling for LDA

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

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

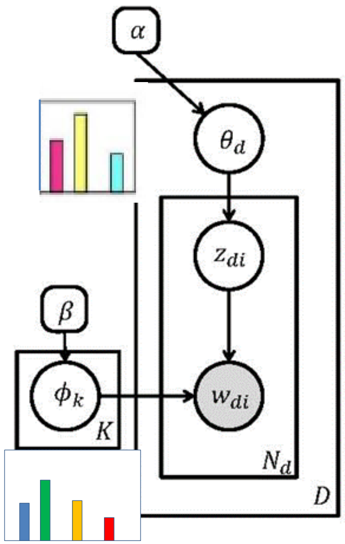
$$= \left\{ \prod_{k=1}^{K} \frac{B(h_\phi(k, \cdot) + \beta)}{B(\beta)} \right\} \left\{ \prod_{d=1}^{D} \frac{B(h_\theta(d, \cdot) + \alpha)}{B(\alpha)} \right\}$$

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

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

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

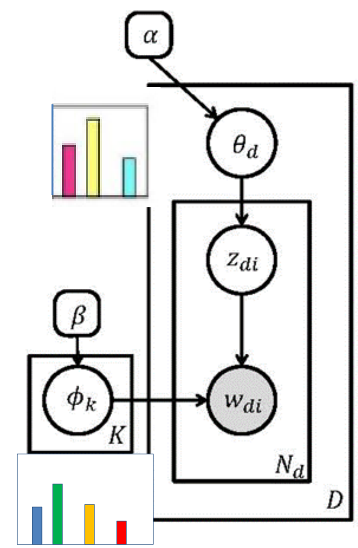
- The Gibbs sampling equation for LDA

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

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

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

Not related with z_{-di}



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

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

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

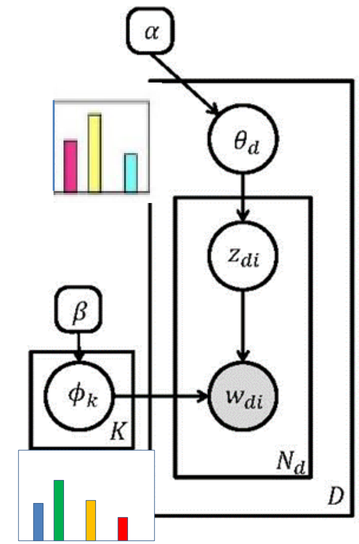
Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA

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

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

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



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

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

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

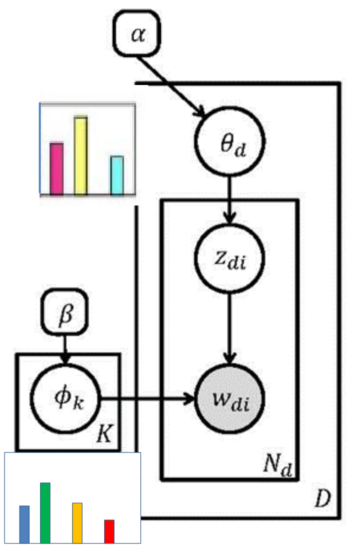
- The Gibbs sampling equation for LDA

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

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

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

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



Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA

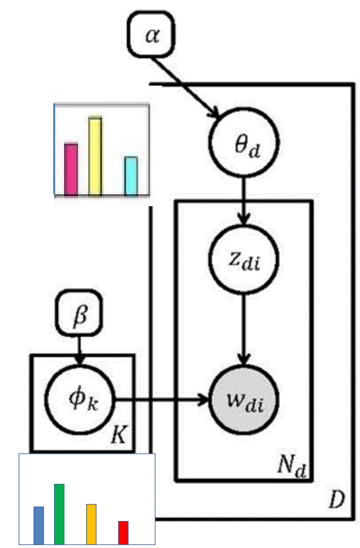
$$p(z_{di}|z_{-di},w,\alpha,\beta) \propto \prod_{k=1}^{K} \frac{B(h_\phi(k,\cdot)+\beta)}{B(h_\phi(k,-di)+\beta)} \times \prod_{d=1}^{D} \frac{B(h_\theta(d,\cdot)+\alpha)}{B(h_\theta(d,-di)+\alpha)}$$

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

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

constant

k depends on the sampled z_{di}



Collapsed Gibbs Sampling for LDA

- The Gibbs sampling equation for LDA

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

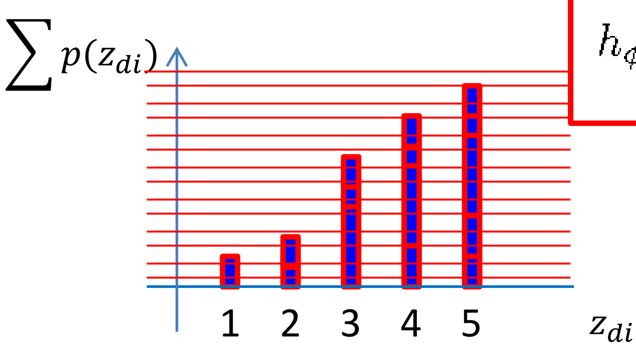
Tractable

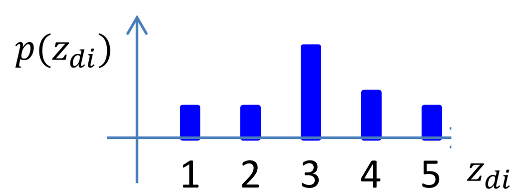
$$\times \prod_{d=1}^{D}\prod_{k=1}^{K} \{h_\theta(d,k) - \delta[z_{di} - k] + \alpha(k)\},$$

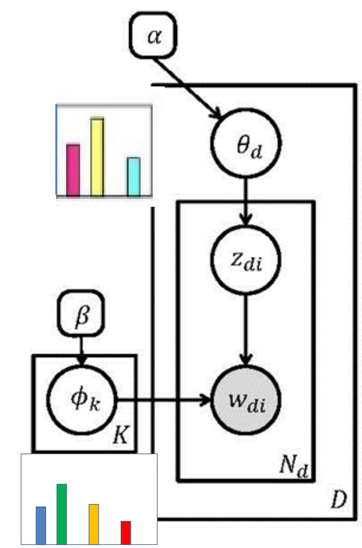
- Sampling

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

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







What is the inferred label of w_{di} ?

Collapsed Gibbs Sampling for LDA

- Estimation of the multinomial parameters θ and ϕ

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

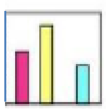
$$= Dir(\theta_d|h_\theta(d,\cdot)+\alpha),$$

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

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

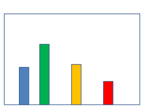
w: 1, 3, 2, 3, 3, 5, 4, 1, 6
z : 1, 2, 2, 2, 1, 1, 2, 1, 2
h_θ(d,2): 5

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 $\leftarrow$ 2
h_ϕ(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

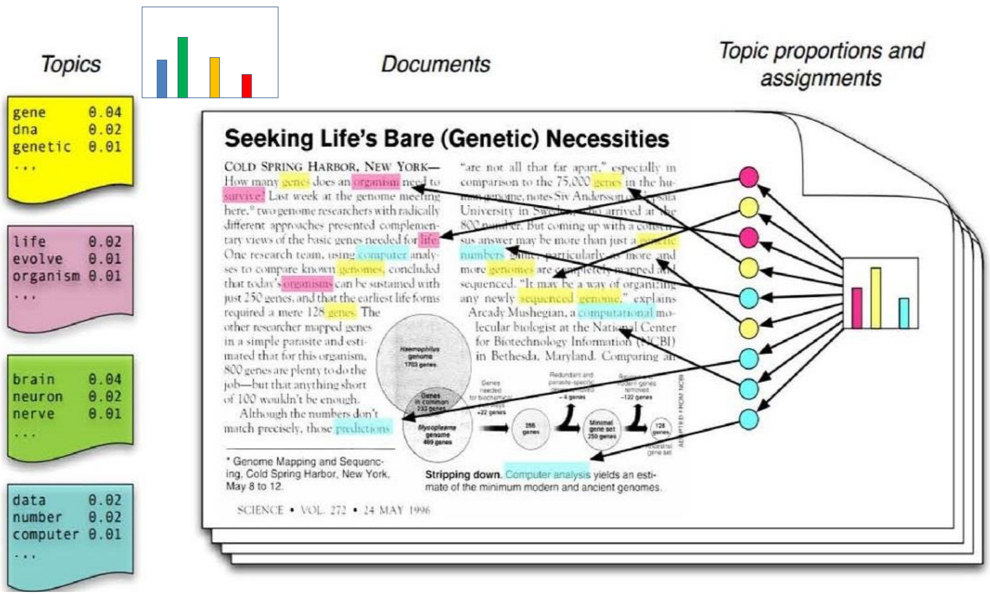
Inference of Bayesian Network: Variational Inference

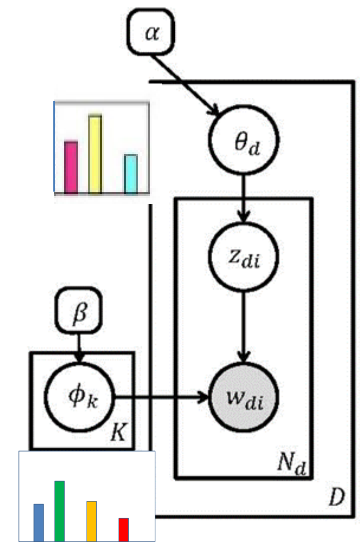
Jin Young Choi

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 toTraffic Pattern Analysis

LDA Model (Topic Modelling)





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

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

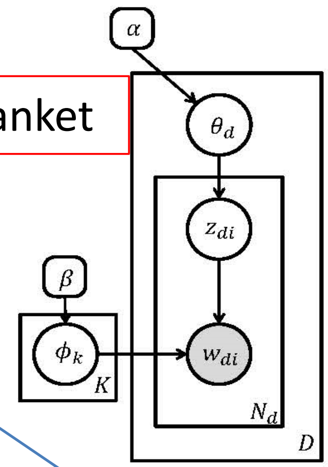
$Likelihood: p(w|z, \theta, \emptyset, \alpha, \beta)$

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

Markov Chain Monte Carlo (MCMC)

- Markov Chain Monte Carlo (MCMC) framework

Markov Blanket



Posteriors

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

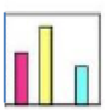
$$= Dir(\theta_d|h_\theta(d,\cdot)+\alpha),$$

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

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

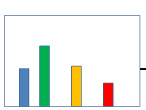
$$h_\theta(d,k) = \sum_{i=1}^{N_d} \delta\left[z_{di}-k\right])$$

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



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

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



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

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

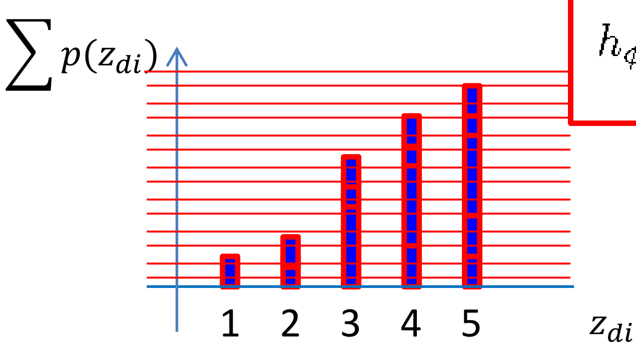
Tractable

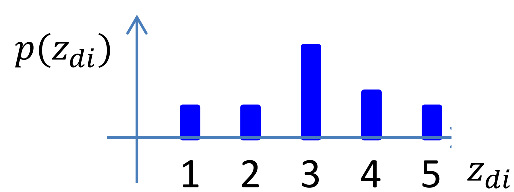
$$\times \prod_{d=1}^{D} \prod_{k=1}^{K} \left\{h_\theta(d,k) - \delta\left[z_{di} - k\right] + \alpha(k)\right\},$$

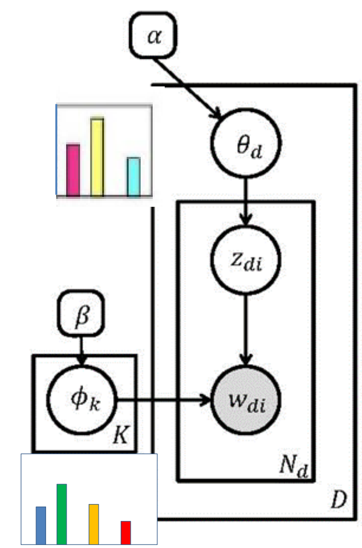
- Sampling

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

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







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)

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

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

$$Likelihood\colon\ p(w|z,\theta,\emptyset,\alpha,\beta)$$

$$Posteriori\colon\ p(z,\theta,\emptyset|w,\alpha,\beta)$$

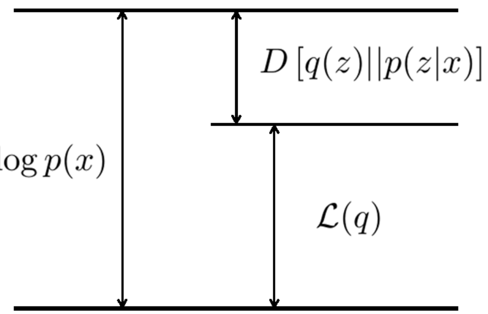
Variational Inference (VI)

- Kullback–Leibler divergence (KL-divergence)

$$p(z|x) = \frac{p(z,x)}{p(x)}$$

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

Minimize $-\mathcal{L}(q) \triangleq \int_z q(z)\log\frac{q(z)}{p(z,x)}dz$; $\log p(x)$: constant

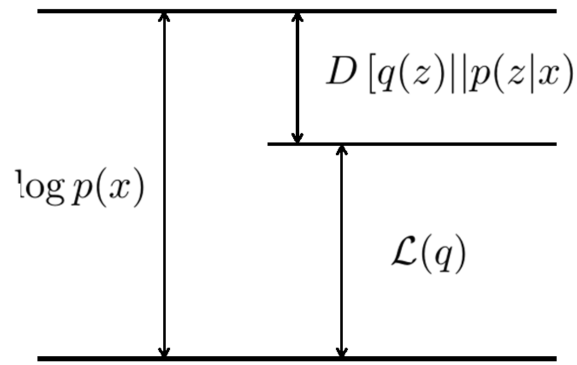


Variational Inference (VI)

- Dual of KL-divergence

$$\begin{aligned}\mathcal{L}(q) &\triangleq \int_z q(z)\log\frac{p(z,x)}{q(z)}dz\\ &= \int_z q(z)\log p(z,x)dz - \int_z q(z)\log q(z)dz\\ &= E_q\left[\log p(z,x)\right] + \underline{H\left[q(z)\right]}.\end{aligned}$$

entropy of $q(z)$



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 $\quad \hat{\lambda} = \max_{\lambda} \bar{\mathcal{L}}(\lambda)$

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

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

$z \rightarrow \phi$
$x \rightarrow w$

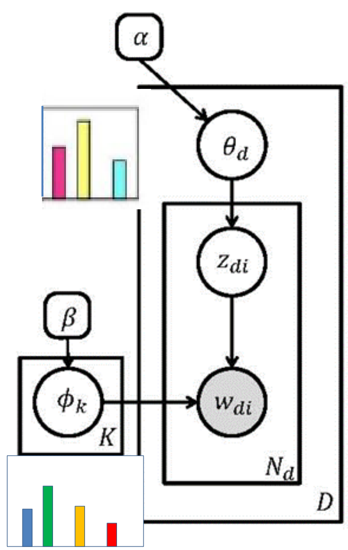
$$p(\phi|w,z) \propto p(w|\phi,z)p(\phi|\alpha) \approx q(\phi|\gamma)$$
$$C\phi_1^{h(1)+\alpha_1}\phi_2^{h(2)+\alpha_2} \ldots \phi_V^{h(V)+\alpha_v} \approx C\phi_1^{\gamma_1}\phi_2^{\gamma_2} \ldots \phi_V^{\gamma_V}$$

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

$z \rightarrow \theta$
$x \rightarrow z$

$$p(\theta|z) \propto p(z|\theta)p(\theta|\alpha) \approx q(\theta|\mu)$$
$$C\theta_1^{h(1)+\alpha_1}\theta_2^{h(2)+\alpha_2} \ldots \theta_K^{h(K)+\alpha_v} \approx C\theta_1^{\mu_1}\theta_2^{\mu_2} \ldots \theta_K^{\mu_K}$$

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



Variational Inference for LDA

- Objective function and joint probability of LDA

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

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

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

Maximization is intractable
→ Variational Inference

where

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

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

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.

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

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

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

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

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

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

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

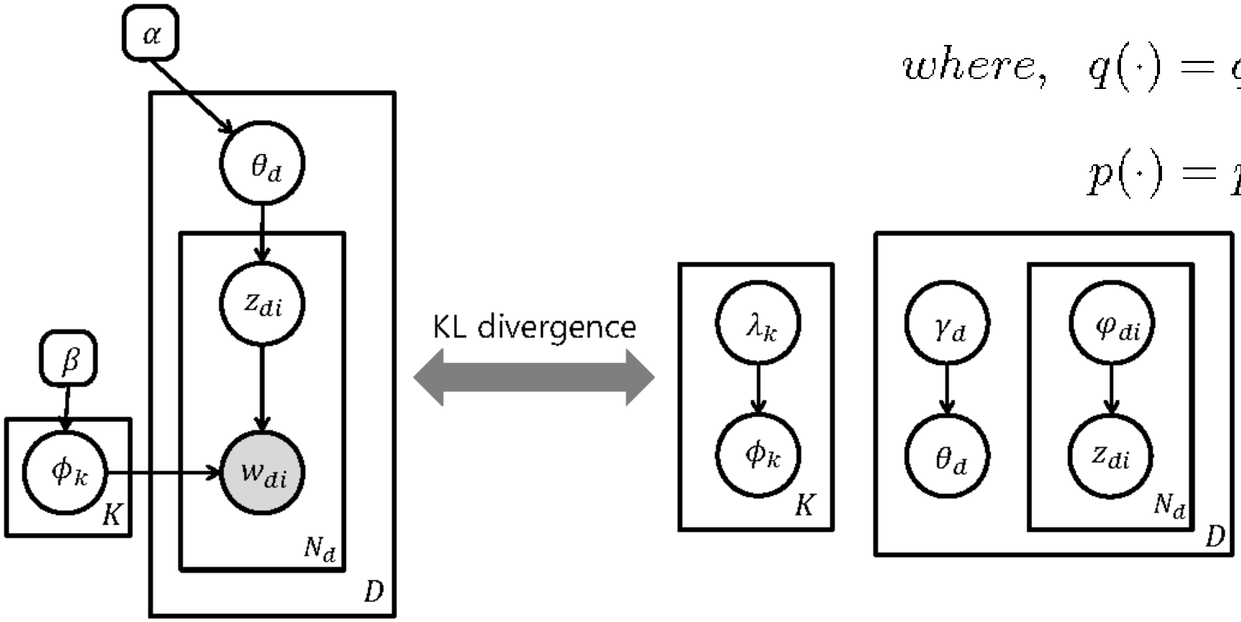
Variational Inference for LDA

- Optimal values of the variational parameters

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

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

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



LDA Model

Approximated Model
Using Variational distribution

$$p(\theta|z) \propto p(z|\theta)p(\theta|\alpha) \approx q(\theta|\gamma)$$
$$C\theta_1^{h(1)+\alpha_1}\theta_2^{h(2)+\alpha_2} \dots \theta_K^{h(K)+\alpha_K} \approx C\theta_1^{\gamma_1}\theta_2^{\gamma_2} \dots \theta_K^{\gamma_K}$$

$$p(\phi|w,z) \propto p(w|\phi,z)p(\phi|\beta) \approx q(\phi|\lambda)$$
$$C\phi_1^{h(1)+\beta_1}\phi_2^{h(2)+\beta_2} \dots \phi_V^{h(V)+\beta_V} \approx C\phi_1^{\lambda_1}\phi_2^{\lambda_2} \dots \phi_V^{\lambda_V}$$

Variational Inference for LDA

- Optimal values of the variational parameters

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

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

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

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

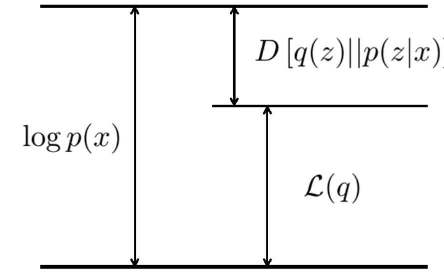
where $$p(\phi,\theta,z|w,\alpha,\beta) = p(\phi,\theta,z,w|\alpha,\beta)/p(w)$$

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

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

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

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

$$\frac{\partial \mathcal{L}(q)}{\partial \varphi_{di}} = 0 \qquad \varphi_{di}(k) \propto \exp\{\underbrace{E_q[\log \theta_d(k)]}_{f(\gamma_d)} + \underbrace{E_q[\log \phi_k(w_{di})]}_{f(\lambda_k)}\}$$

$$\frac{\partial \mathcal{L}(q)}{\partial \gamma_d} = 0 \qquad \gamma_d(k) = \alpha(k) + \sum_{i=1}^{N_d} \varphi_{di}(k)$$

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

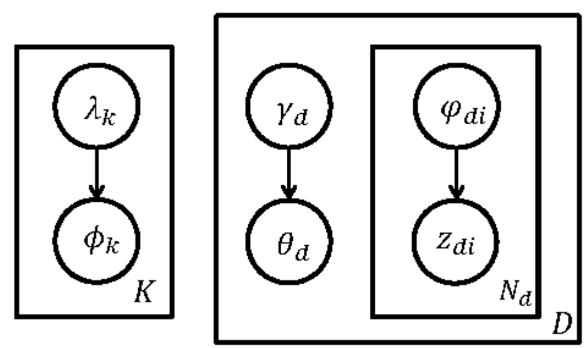
$$h_\theta(d,k) = \sum_{i=1}^{N_d} \delta\left[z_{di} - k\right])$$

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

$\phi_k|\lambda_k \sim Dirichlet(\phi_k|\lambda_k)$ → $Dir(\phi_k|h_\phi(k,\cdot) + \beta)$.

$\theta_d|\gamma_d \sim Dirichlet(\theta_d|\gamma_d)$ → $Dir(\theta_d|h_\theta(d,\cdot) + \alpha)$

$z_{di}|\varphi_{di} \sim Multi(z_{di}|\varphi_{di})$. → $Multi(z_{di}|\theta_d,, w_{di})$ –

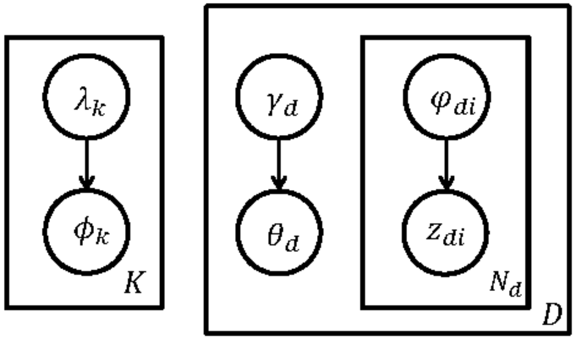


Variational Inference for LDA

- Expectation

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

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



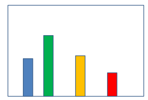
where $\Psi(.)$ is the digamma function (Blei et al. 2003) given by

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

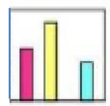
- The iterative updates of the variational parameters are guaranteed to converge into a stationary point
- For the iteration, φ are updated with λ and γ fixed, and λ and γ are updated given the fixed φ

Variational Inference for LDA

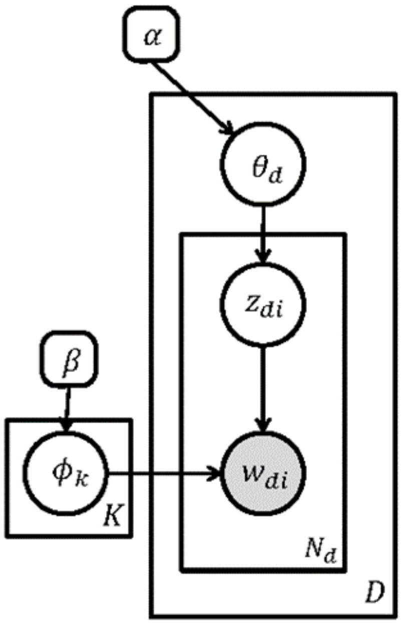
- The final distribution results ϕ and θ



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

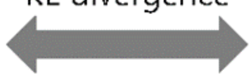


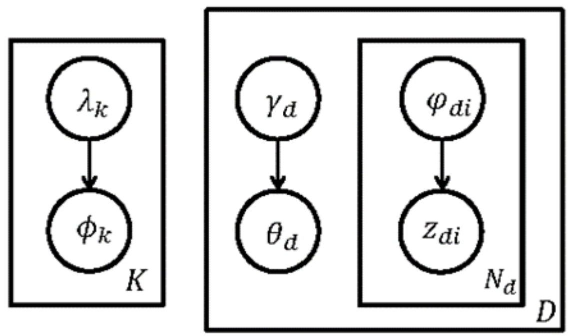
$$\hat{\theta}_d(k) = E_{\underbrace{q(\theta_d|\gamma_d)}_{Dirichlet}}[\theta_d(k)|\gamma_d] = \frac{\gamma_d(k)}{\sum_{k=1}^{K}\gamma_d(k)}.$$

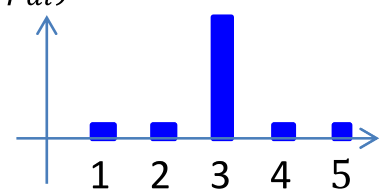


LDA Model

KL divergence







Approximated Model
Using Variational distribution

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

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

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

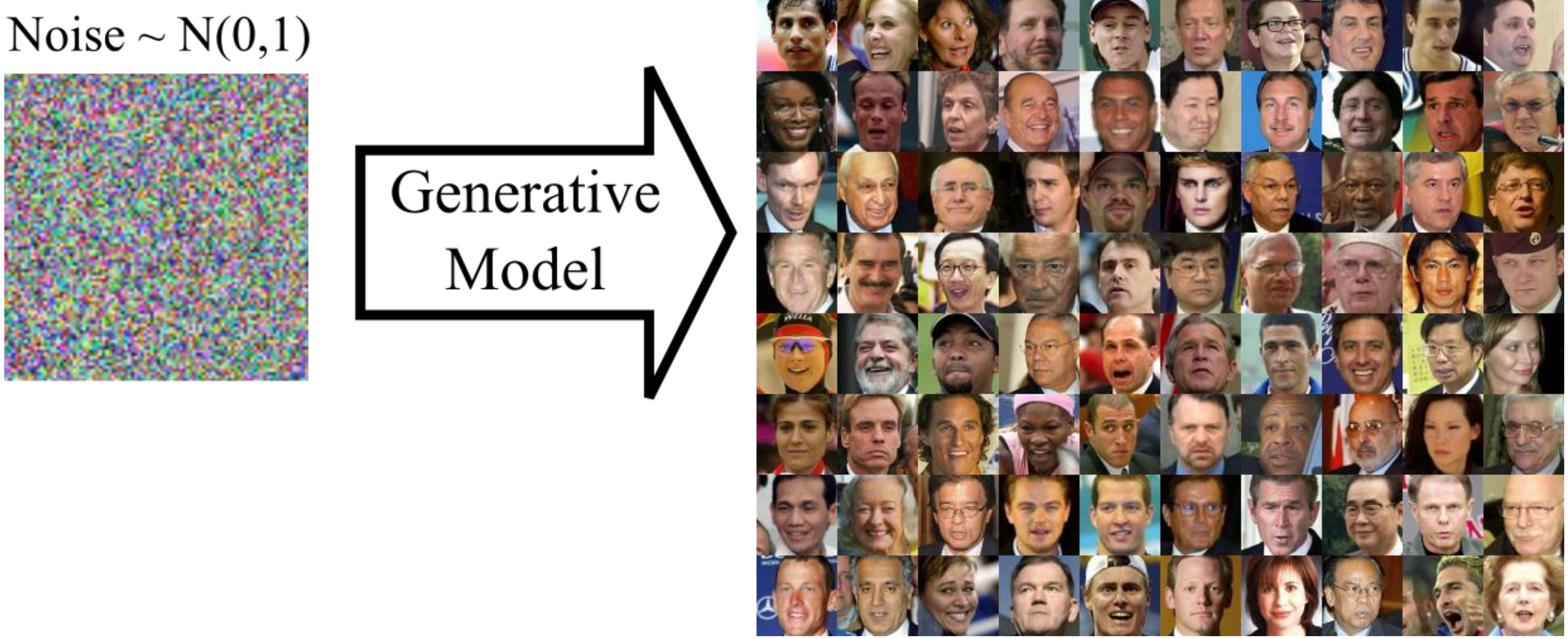
References for LDA and VI

- 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

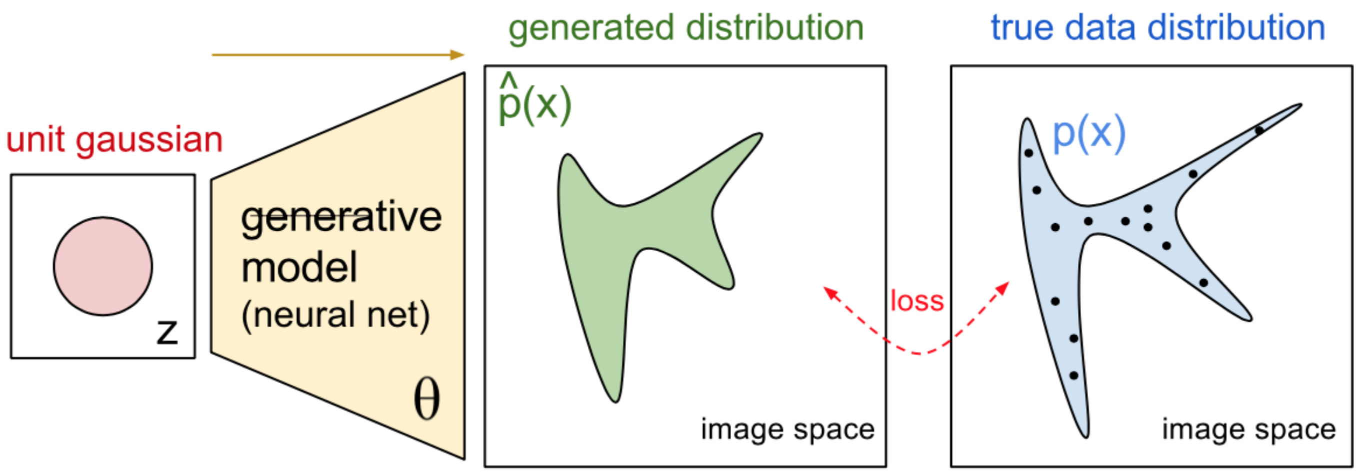
Ex) Image Generation



Generative Image Modeling

Goal: Model the distribution p(x)

- cf) Discriminative approach

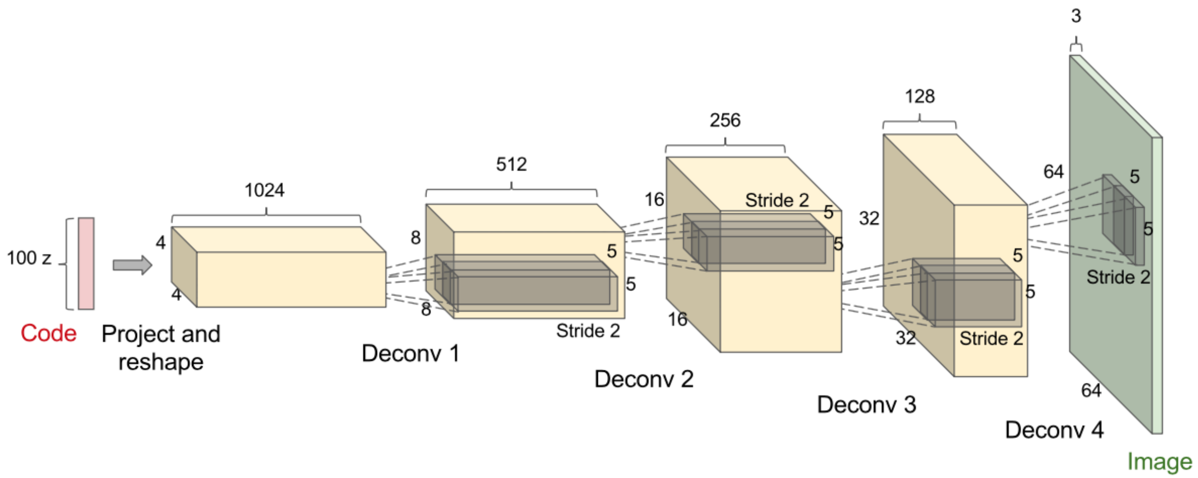


Issue: High Dimensional
low resolution (64x64x3 = 12,288)

Generative Image Modeling

Goal: Model the distribution p(x)

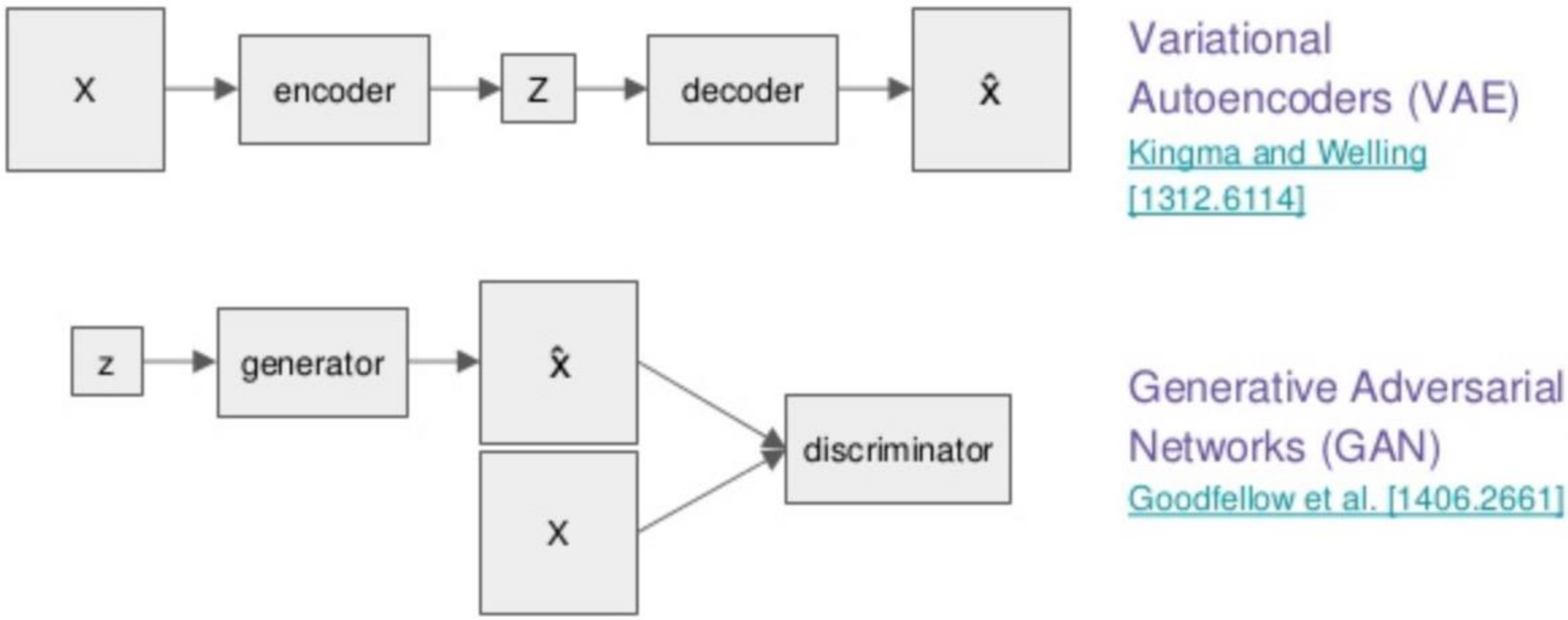
- Using Deep Neural Networks (CNNs)



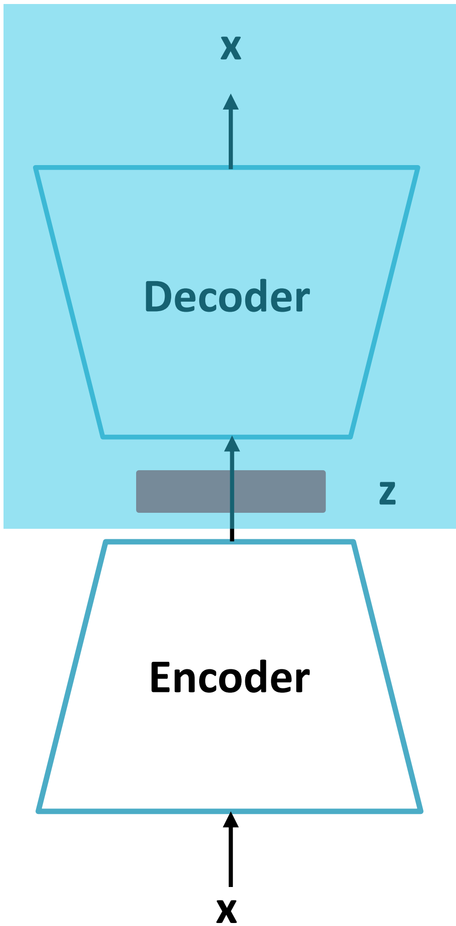
Deep Generative Models

Two prominent approaches

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



Variational Auto-encoder (VAE)

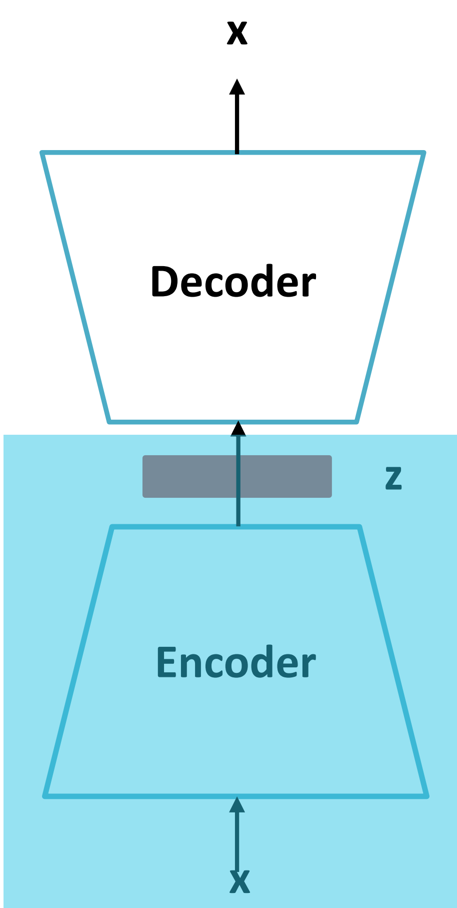


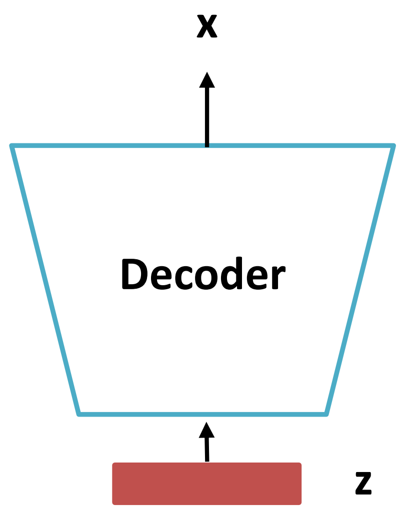
Reconstruction Loss   Variational Inference

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

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

Variational Auto-encoder (VAE)





$Sampling\ z\ from\ N(0, I)$

Variational Inference

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

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

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