Gibbs Sampling

Subeen Cha
Machine Learning Laboratory
Department of Statistics
EWHA Womans University

0. Introduction

- Sometimes, exact inference is intractable. → can use approximation
  - ex. Numerical sampling methods(Monte Carlo, etc)
- Sampling
  - General Idea: Obtain a set of samples $\boldsymbol{z}^{(l)}$ $(l = 1, \cdots, L)$ drawn independently from the distribution $p(\boldsymbol{z})$.
  - Then, the expectation can be approximated by a finite sum.

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z} = \frac{1}{L}\sum_{l=1}^{L} f\left(z^{(l)}\right)$$

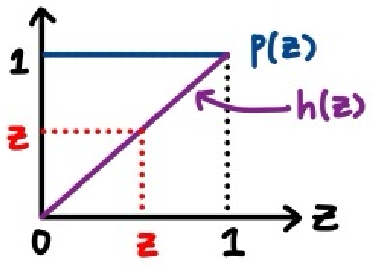
  - Some problems
    - The samples $\{\boldsymbol{z}^{(l)}\}$ might not be independent.
    - The expectation may be dominated by regions of small probability.
      - $f(\boldsymbol{z})$: small vs $p(\boldsymbol{z})$: large

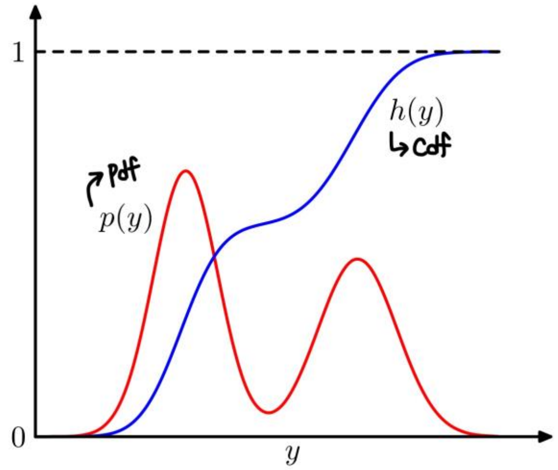
1. Basic Sampling Algorithms

1) Standard distributions

- Suppose that z is uniformly distributed over the interval $(0, 1)$, and that we transform the values of z using some function $f(\cdot)$. $\Rightarrow$ $y = f(z)$

$z \sim U(0,1)$, $y = f(z) \longrightarrow z = f^{-1}(y)$





Jacobi transformation

$$h(y) \equiv F_Y(y) = P(Y \le y) = \int_{-\infty}^{y} p(\hat{y}) \cdot d\hat{y} = P(f(z) \le y) = P(z \le f^{-1}(y)) = F_Z(f^{-1}(y)) = h(z) = z \quad (11.6)$$

$$p(y) \equiv f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_Z(f^{-1}(y)) = f_Z(\underbrace{f^{-1}(y)}_{z}) \cdot \frac{d}{dy} \underbrace{f^{-1}(y)}_{z} = f_Z(z) \frac{dz}{dy} = p(z)\left|\frac{dz}{dy}\right| \quad (11.5)$$

$\rightarrow$ transform z using the inverse of the indefinite integral of the desired distribution $p(y)$

$$y = h^{-1}(z)$$

1) Standard distributions

- Ex 1. exponential distribution
  - pdf: $\mathrm{p}(\mathrm{y}) = \lambda \exp(-\lambda \mathrm{y}), \quad 0 \leq \mathrm{y} < \infty$
  - cdf: $z = h(y) = \int_0^y p(\hat{y}) d\hat{y}$

$$= 1 - \exp(-\lambda y)$$

$$\rightarrow \exp(-\lambda y) = 1 - z$$

$$\rightarrow -\lambda y = \ln(1 - z) \quad (\because 0 < z \leq 1)$$

$$\rightarrow y = -\lambda^{-1} \ln(1 - z) \sim \mathrm{Exp}(\lambda)$$

- Ex 2. Cauchy distribution

$$p(y) = \frac{1}{\pi} \cdot \frac{1}{1+y^2}$$

$$h(y) = \int_{-\infty}^{y} \frac{1}{\pi} \cdot \frac{1}{1+\hat{y}^2} d\hat{y} = \int_{-\frac{\pi}{2}}^{\tan^{-1}(y)} \frac{1}{\pi} \cdot \cancel{\frac{1}{1+\tan^2\theta}} \cdot \cancel{\sec^2\theta}\, d\theta$$

$\hat{y} = \tan\theta$

$d\hat{y} = \sec^2\theta \cdot d\theta$

$$= \left[\frac{1}{\pi} \cdot \theta \right]_{-\frac{\pi}{2}}^{\tan^{-1}(y)} = \frac{1}{\pi} \cdot \left\{ \tan^{-1}(y) + \frac{\pi}{2} \right\} = z$$

$$\rightarrow \tan^{-1}(y) = \pi z - \frac{\pi}{2}$$

$$\rightarrow y = \tan\left[\pi\left(z - \frac{1}{2}\right)\right]$$

  - multiple variables → Use Jacobian

$$\mathrm{p}(\mathrm{y}_1, \cdots, \mathrm{y}_\mathrm{M}) = \mathrm{p}(\mathrm{z}_1, \cdots, \mathrm{z}_\mathrm{M}) \left| \frac{\partial(z_1, \cdots, z_M)}{\partial(y_1, \cdots, y_M)} \right|$$

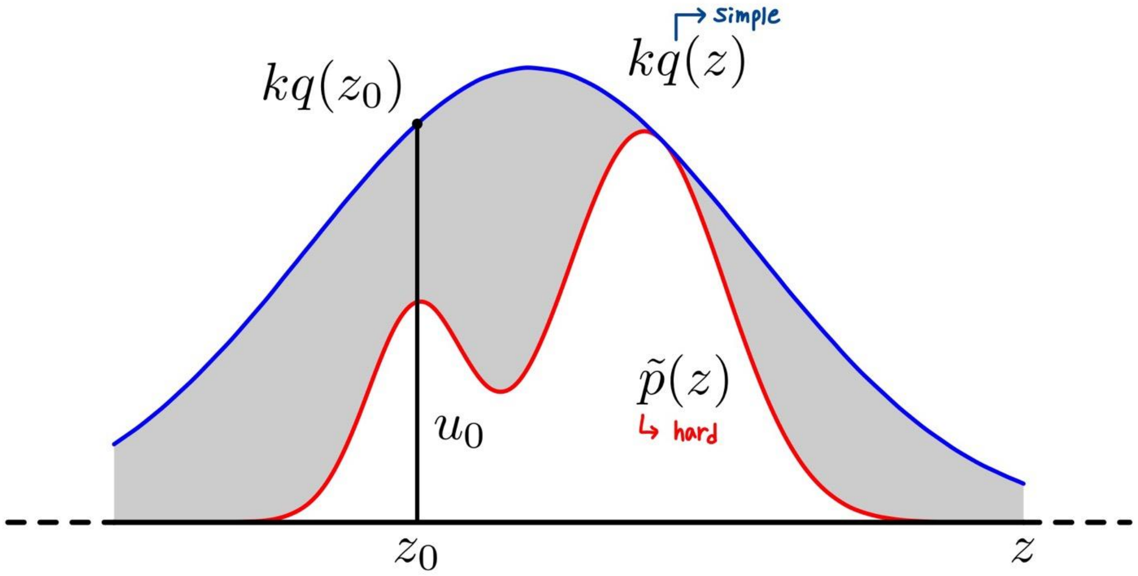
- Generally, calculating and then inverting the indefinite integral of the required distribution is intractable.
  - More general strategy is required: rejection sampling and importance sampling

2) Rejection sampling

- Suppose that sampling directly from $p(\boldsymbol{z})$ is difficult, but easily able to evaluate $p(\boldsymbol{z})$ for any given value of $\boldsymbol{z}$: $p(z) = \frac{1}{Z_p}\tilde{p}(z)$, Z_p: unknown
- proposal distribution($q(z)$)
  - simpler distribution that we can readily draw samples
  - Choose a constant k s.t $kq(z) \geq \tilde{p}(z)$ for all values of z.
    - $kq(z)$: comparison function
- Each step of the rejection sampler involves generating two random numbers.
  - $z_0 \sim q(z)$
  - $u_0 \sim U[0, kq(z_0)]$
  - If $u_0 > \tilde{p}(z_0)$
    - reject z_0 and resample(grey shaded region)
  - Else
    - Accept z_0
  - Samples are accepted with probability $\tilde{p}(z)/kq(z)$.

$$p(\text{accept}) = \int \{\tilde{p}(z)/kq(z)\}q(z)dz = \frac{1}{k}\int \tilde{p}(z)dz$$

$\Rightarrow$ choose k as small as possible



2-α) Adaptive Rejection sampling

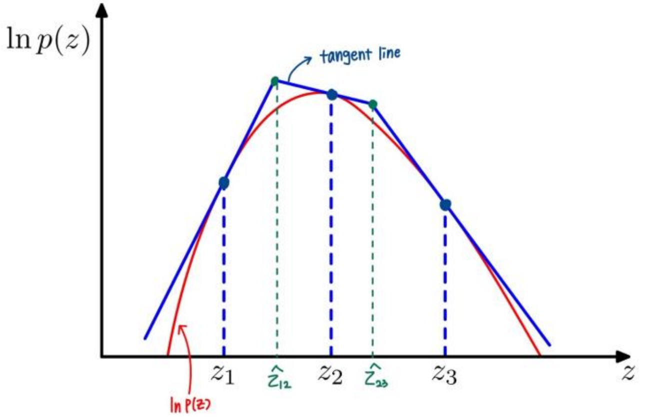
- Construction of an envelope function($kq(z)$) is particularly easy if $p(z)$ is log-concave.
  - Choose some initial set of grid points: $\mathrm{z}_1, \mathrm{z}_2, \mathrm{z}_3$
  - Calculate $\ln p\,(z_i)$ and gradient(λ_i).
  - Construct an envelope function($q(z)$).
    - set of piecewise exponential distributions

$$q(z) = k_i \lambda_i \exp\{-\lambda_i(z - z_i)\}, \quad \widehat{z_{i-1,i}} < z \le \widehat{z_{i,i+1}},$$

k_i: gradient of the tangent line at z_i

- Once a sample has been drawn, apply the usual rejection criterion.
  - If $u_0 \le \tilde{p}(z_0)$
    - Accept z_0
  - Else
    - Reject z_0. $\rightarrow$ Incorporate into the set of grid points.
    - Compute a new tangent line. $\rightarrow$ update envelope function $\Rightarrow$ adaptive

$\Rightarrow$ Envelope function becomes a better approximation of the desired distribution $p(z)$.



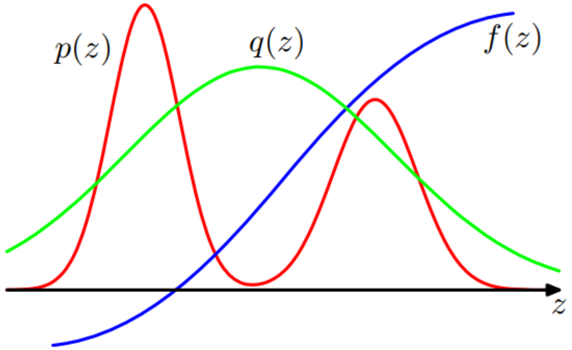
2) Rejection sampling_limitation

- For high-dimensional problems, the acceptance rate decreases exponentially as the number of dimensions increases.
- Ex> multivariate Gaussian distribution
  - proposal distribution($kq(z)$): $N(0, \sigma_p^2 \mathrm{I})$
  - In D-dimensions the optimum value of k is given by $k = \left(\sigma_q / \sigma_p\right)^D$.
  - If the proposal distribution's variance is only slightly larger than the target distribution's, the acceptance ratio can become extremely low.
    - about 1/20,000 for D = 1,000 dimensions

- Finding a good proposal distribution in high-dimensional problem is difficult.
  - Alternatives) importance sampling, etc

3) Importance sampling(= weighted sampling)

- Main goal of sampling
  - Generate the exact pdf. (X)
  - Calculating the expectation of pdf or calculating a certain probability (O)



- Uniform sampling
  - discretize $\mathbf{z}$-space into a uniform grid and evaluate partial sums

    $\mathbb{E}[f] \simeq \sum_{l=1}^{L} f\left(z^{(l)}\right) p\left(z^{(l)}\right)$
  - But the number of terms in the summation grows exponentially with the dimensionality of $\mathbf{z}$.
  - In addition, most probability distributions have most masses inside relatively small regions in $\mathbf{z}$ space in high-dimensional problems.
    - Only a very small proportion of the samples will make a significant contribution to the sum.

⇒ importance sampling

3) Importance sampling

- Again, use proposal distribution $q(z)$ from which we can easy draw samples.
  - Sampling from $q(z)$: $\{z^{(l)}\}$
  - Then, calculate the expectation.

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z} \simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p\left(\mathbf{z}^{(l)}\right)}{q\left(\mathbf{z}^{(l)}\right)}f\left(\mathbf{z}^{(l)}\right)$$

    - $\boldsymbol{L}$: # of samples, $\mathbf{z}^{(l)}$: sample of Z

  - Importance weight: $r_l = p\left(z^{(l)}\right)/q\left(z^{(l)}\right)$
    - kind of weighted average
    - used to correct the bias introduced by sampling from the wrong distribution($q(z)$)

1. Basic Sampling Algorithms

3) Importance sampling

- Normalization
  - case that the distribution $p(\boldsymbol{z})$ can only be evaluated up to a normalization constant
    - i.e., $p(\boldsymbol{z}) = \tilde{p}(\boldsymbol{z})/Z_p$ where $\tilde{p}(\boldsymbol{z})$ can be evaluated easily, whereas Z_p is unknown
  - can use $q(\boldsymbol{z}) = \tilde{q}(\boldsymbol{z})/Z_q$

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

$$= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z} = \int f(\mathbf{z})\frac{\tilde{p}(\mathbf{z})/Z_p}{\tilde{q}(\mathbf{z})/Z_q}q(\mathbf{z})d\mathbf{z}$$

$$= \frac{Z_q}{Z_p}\int f(\mathbf{z})\underline{\frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}}q(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

$$\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\underline{\widetilde{r}_l f(\mathbf{z}^{(l)})}.$$

$$\boxed{\begin{aligned}&\mathbb{E}[f] \simeq \textstyle\sum_{l=1}^{L} w_l f(\mathbf{z}^{(l)})\\ &\text{where}\quad w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^{(l)})/q(z^{(l)})}{\sum_m \tilde{p}(z^{(m)})/q(z^{(m)})}\end{aligned}}$$

- $\tilde{r}_l = \dfrac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}$
- $\dfrac{Z_p}{Z_q} = \dfrac{1}{Z_q}\int \tilde{p}(z)dz = \int \dfrac{\tilde{p}(z)}{\tilde{q}(z)}\tilde{q}(z)dz \simeq \dfrac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l$

3) Importance sampling_likelihood weighted sampling

- Significant difference from $p(z)$ and $q(z)$ can reduce the efficiency of sampling.
  - improvement → likelihood weighted sampling

- Likelihood weighted sampling
  - based on ancestral sampling of the variables
    - make one pass through the set of variables in the order $z_1, \cdots, z_M$ sampling from the conditional distributions $p(\boldsymbol{z}_i|\mathrm{pa}_i)$.
  - Check each variable in turn.
    - If that variable is in the evidence set(= observed variables), then it is just set to its instantiated value.
    - Else, sampling from the conditional distribution $p(\boldsymbol{z}_i|\mathrm{pa}_i)$
  - Weights

$$r(\mathbf{z}) = \prod_{\mathbf{z}_i \notin \mathbf{e}} \frac{p(\mathbf{z}_i|\mathrm{pa}_i)}{p(\mathbf{z}_i|\mathrm{pa}_i)} \prod_{\mathbf{z}_i \in \mathbf{e}} \frac{p(\mathbf{z}_i|\mathrm{pa}_i)}{1} = \prod_{\mathbf{z}_i \in \mathbf{e}} p(\mathbf{z}_i|\mathrm{pa}_i)$$

3-α) Sampling Importance Resampling

Improves weight balance problem in importance sampling

- Process
  - 1. Sampling: draw L samples $\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(L)}$ from $q(\boldsymbol{z})$.
  - 2. Weight construction: $w_l = \frac{\widetilde{r_l}}{\sum_m \widetilde{r_m}} = \frac{\tilde{p}\left(z^{(l)}\right)/q\left(z^{(l)}\right)}{\sum_m \tilde{p}\left(z^{(m)}\right)/q\left(z^{(m)}\right)}$
  - 3. Resampling: resample # L samples from the sample sets ($\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(L)}$).
    - Sampling probabilities are given by the weights $(w_1, \cdots, w_L)$.
- The approximation improves as the sampling distribution $q(z)$ gets closer to the desired distribution $p(z)$.
  - If $L \rightarrow \infty$, the distribution becomes correct.
    - When $q(z) = p(z)$, the samples sets($\boldsymbol{z}^{(1)}, \cdots, \boldsymbol{z}^{(L)}$) have the desired distribution($p(z)$) , and the weights become uniformly $w_n = 1/L$.
    - So, the resampled values also have the desired distribution.

Sampling and the EM Algorithm

- M-step
  - optimize Q(log-likelihood estimates) w.r.t θ(model parameters).
  - Update the model parameter θ given the value z.

$$Q\left(\theta, \theta^{old}\right) = \int p\left(\mathbf{Z} \middle| \mathbf{X}, \theta^{old}\right) \ln p(\mathbf{Z}, \mathbf{X}|\theta) d\mathbf{Z}$$

- E-step
  - Update z given the model parameter value θ.
  - can use sampling methods to approximate the integral by a finite sum over samples $\left\{Z^{(l)}\right\}$

$$Q\left(\theta, \theta^{old}\right) \simeq \frac{1}{L} \sum_{l=1}^{L} ln\, p\left(\mathbf{Z}^{(l)}, \mathbf{X} \middle| \theta\right)$$

2. Markov Chain Monte Carlo(MCMC)

2-1. Markov Chain

- A Markov chain(or Markov process) is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. Ⓒ Wikipedia
- A series of random variables $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(M)}$ s.t the following conditional independence property holds
  - Conditional independence: $p\left(\mathbf{z}^{(m+1)} \middle| \mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(m)}\right) = p\left(\mathbf{z}^{(m+1)} \middle| \mathbf{z}^{(m)}\right), \quad m \in \{1, \cdots, M-1\}$

2-1. Markov Chain

- Terminologies & Properties
  - transition probabilities
    - We can specify the Markov chain by giving the probability distribution for the initial variable $p(\mathbf{z}^{(0)})$ together with the conditional probabilities for subsequent variables.
      $T_m\left(z^{(m)}, z^{(m+1)}\right) \equiv p\left(z^{(m+1)} \middle| z^{(m)}\right)$
    - A Markov chain is called **homogeneous** if the transition probabilities are the same for all m.
  - invariant(= stationary)
    - A distribution is said to be invariant(or stationary) if each step in the chain leaves that distribution invariant.
    - For a homogeneous Markov chain with transition probabilities $T(z', z)$, the distribution $p^*(z)$ is invariant if $p^*(z) = \sum_{z'} T(z', z) p^*(z')$.

2-1. Markov Chain

- Terminologies & Properties
  - detailed balance condition
    $p^*(\mathbf{z})T(\mathbf{z},\mathbf{z}') = p^*(\mathbf{z}')T(\mathbf{z}',\mathbf{z})$
    - Transition probability that satisfies detailed balance w.r.t a particular distribution will leave that distribution invariant. $$\sum_{\mathbf{z}'} p^\star(\mathbf{z}')T(\mathbf{z}',\mathbf{z}) = \sum_{\mathbf{z}'} p^\star(\mathbf{z})T(\mathbf{z},\mathbf{z}') = p^\star(\mathbf{z}) \sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) = p^\star(\mathbf{z}).$$
    - A Markov chain that respects detailed balance is said to be **reversible**.
  - Ergodicity
    - For $m \to \infty$, the distribution $p(\mathbf{z}^{(m)})$ converges to the required invariant distribution $p^*(\mathbf{z})$, irrespective of the choice of initial distribution $p(\mathbf{z}^{(0)})$.
    - $p^*(\mathbf{z})$: equilibrium distribution

2-1. Markov Chain

- Markov Chain for Sampling
  - In previous, we do not use the past records.
    - Every sampling is independent.
  - MCMC generates continuous samples.
    - $\{z^{(1)}, z^{(2)}, \cdots\}$: forms a Markov Chain
    - Also, maintain a record of the current state $z^{(\tau)}$ and the proposal distribution $q(z|z^{(\tau)})$.
    - At each cycle of the algorithm, we generate a candidate sample $\mathbf{z}^*$ from the proposal distribution and then accept the sample according to an appropriate criterion.

2-2. The Metropolis-Hastings Algorithm

- Metropolis Algorithm

  General algorithm of MCMC

  - Assumption) The proposal distribution is symmetric: $q(\mathbf{z}_A|\mathbf{z}_B) = q(\mathbf{z}_B|\mathbf{z}_A)$
  - Process
    - Current value: $z^{(\tau)}$
    - Propose a candidate $z^* \sim q\left(z^*\middle|z^{(\tau)}\right)$. $q(\,)$: proposal distribution
    - With an acceptance probability $A\left(\mathbf{z}^*, \mathbf{z}^{(\tau)}\right) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)}{\tilde{p}\left(\mathbf{z}^{(\tau)}\right)}\right)$,
      - Choose a random number u with uniform distribution over the unit interval $(0, 1)$.
      - If $A\left(\mathbf{z}^*, \mathbf{z}^{(\tau)}\right) > u$: accept, $\mathbf{z}^{(\tau+1)} \rightarrow \mathbf{z}^*$
      - Else: reject, $\mathbf{z}^{(\tau+1)} \rightarrow \mathbf{z}^{(\tau)}$
        - The candidate point $\mathbf{z}^*$ is discarded. + The previous sample is included instead in the final list of samples.
        - Another candidate sample is drawn from the distribution.

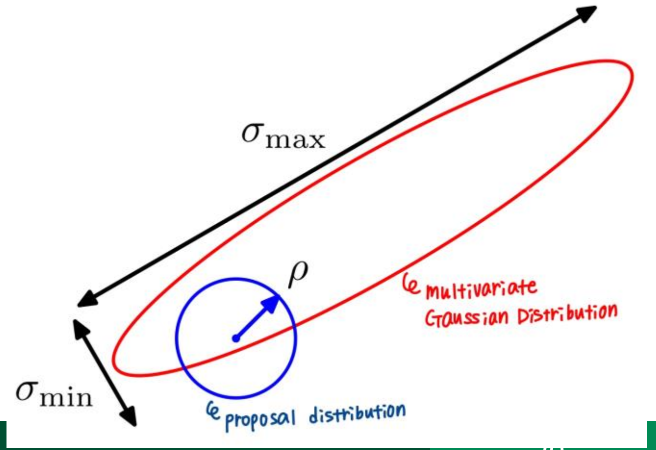
2-2. The Metropolis-Hastings Algorithm

- Metropolis-Hastings Algorithm
  - Remove the symmetric assumption. → more general case
  - Now, acceptance probability becomes $A\left(\mathbf{z}^*, \mathbf{z}^{(\tau)}\right) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)q_k\left(\mathbf{z}^{(\tau)}\middle|\mathbf{z}^*\right)}{\tilde{p}\left(\mathbf{z}^{(\tau)}\right)q_k\left(\mathbf{z}^*\middle|\mathbf{z}^{(\tau)}\right)}\right)$.
  - $p(\mathbf{z})$ is an invariant distribution of the Markov chain defined by the Metropolis-Hastings algorithm.
    - Pf> Show that detailed balance condition holds.
      - Detailed balance condition: $p^*(\mathbf{z})T(\mathbf{z}, \mathbf{z}') = p^*(\mathbf{z}')T(\mathbf{z}', \mathbf{z})$

$$\begin{aligned} p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z})A_k(\mathbf{z}', \mathbf{z}) &= \min\left(p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z}), p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}')\right) \\ &= \min\left(p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}'), p(\mathbf{z})q_k(\mathbf{z}'|\mathbf{z})\right) \\ &= p(\mathbf{z}')q_k(\mathbf{z}|\mathbf{z}')A_k(\mathbf{z}, \mathbf{z}') \end{aligned}$$

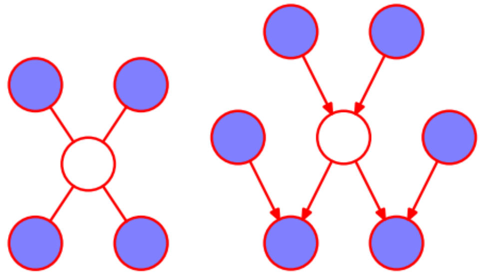
2-2. The Metropolis-Hastings Algorithm

- Metropolis-Hastings Algorithm
  - The specific choice of proposal distribution can have a marked effect on the performance of the algorithm.
    - continuous state spaces: Gaussian centered on the current state
  - There is an important trade-off in determining the variance parameter.
    - Small variance: high transition approval rate, but slow progress through the state space
    - Large variance: high rejection rate
  - The size of the proposed distribution(ρ) should be of the same order as the smallest length scale($\sigma_{\min}$).
    - explores the distribution along the more extended direction by means of a random walk
    - can have very slow convergence if the distributions vary are very different in different directions



Intro

- Gibbs Sampling: special case of the Metropolis-Hastings algorithm
- Consider a Metropolis-Hastings sampling step involving the variable z_k.
    - $z^{(\tau)} = \left(z_k^{(\tau)}, z_{\backslash \mathrm{k}}^{(\tau)}\right) \to z^* = \left(z_k^*, z_{\backslash \mathrm{k}}^*\right)$
    - the remaining variables $\mathbf{z}_{\backslash \mathrm{k}}$ remain fixed: $z_{\backslash k}^* = \mathbf{z}_{\backslash k}$
    - transition probability: $q_k(\mathbf{z}^*|\mathbf{z}) = p\left(z^*\middle|\mathbf{z}_{\backslash k}\right)$
    - Then, the acceptance probability becomes 1. → always accepted
        - $p(\mathbf{z}) = p\left(z_k\middle|\mathbf{z}_{\backslash k}\right)p\left(\mathbf{z}_{\backslash k}\right)$ → $A(\mathbf{z}^\star, \mathbf{z}) = \frac{p(\mathbf{z}^\star)q_k(\mathbf{z}|\mathbf{z}^\star)}{p(\mathbf{z})q_k(\mathbf{z}^\star|\mathbf{z})} = \frac{p(z_k^\star|\mathbf{z}_{\backslash k}^\star)p(\mathbf{z}_{\backslash k}^\star)p(z_k|\mathbf{z}_{\backslash k}^\star)}{p(z_k|\mathbf{z}_{\backslash k})p(\mathbf{z}_{\backslash k})p(z_k^\star|\mathbf{z}_{\backslash k})} = 1$
    - Applicability
        - depends on the ease with which samples can be drawn from the conditional distributions $p\left(z_k\middle|\mathbf{z}_{\backslash k}\right)$
        - Can simplify using the Markov blanket



Concept of Gibbs Sampling

- Each step involves replacing the value of one of the variables by a value drawn from the distribution of that variable conditioned on the values of the remaining variables.
- Repeated either by cycling through the variables in some particular order or by choosing the variable to be updated at each step at random from some distribution
- Process

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

4. Slice Sampling

Intro

- Metropolis algorithm is sensitive to the step size.
  - too small: slow decorrelation due to random walk behavior
  - too large: inefficiency due to a high rejection rate

  ⇒ Slice Sampling provides an adaptive step size that is automatically adjusted to match the characteristics of the distribution.

- Univariate case
  - Introduce an auxiliary variable u and draw samples from the joint (z,u) space.
  - Goal: sample uniformly from the area under the distribution
    - joint distribution: $$\widehat{p}(z,u) = \begin{cases} 1/Z_p & \text{if } 0 \leqslant u \leqslant \widetilde{p}(z) \\ 0 & \text{otherwise} \end{cases}$$
    - marginal distribution over z: $$p(z) = \int \tilde{p}(z,u)\,du = \tilde{p}(z)/Z_p$$

Methology

- Procedure
  - Alternately sample z and u.
    - Given the current value $z^{(\tau)}$, select u uniformly in the range $0 \le \mathrm{u} \le \tilde{\mathrm{p}}(\mathrm{z})$
      - creates a "slice" through the distribution
    - Then, fix u and sample z from the 'slice' through the distribution defined by $\{\mathrm{z}\colon \tilde{\mathrm{p}}(\mathrm{z}) > \mathrm{u}\}$
  - Adjust the region based on the characteristic length scales of the distribution
  - Repeat until a valid z is found.

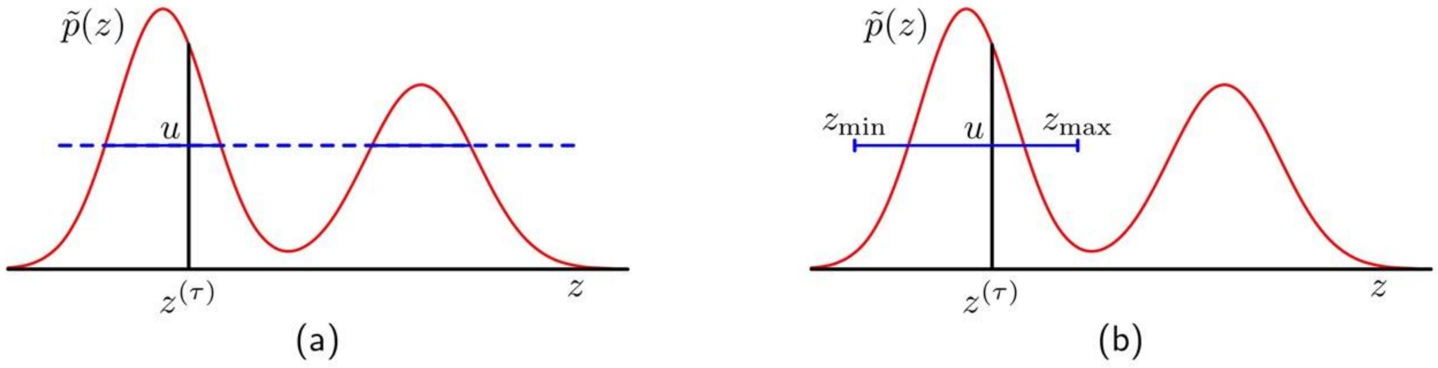


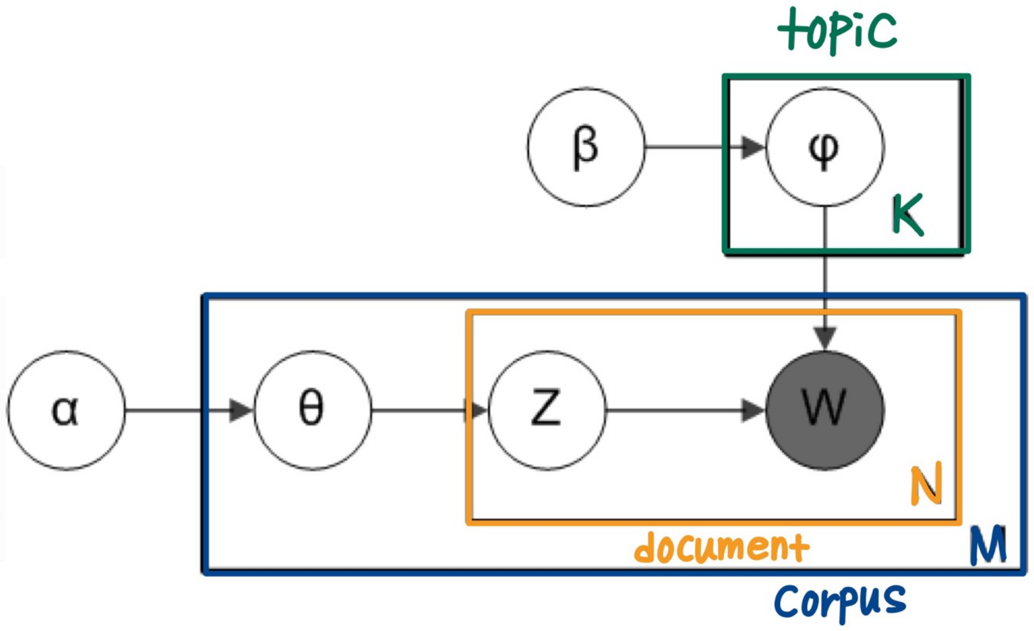
Figure 11.13 Illustration of slice sampling. (a) For a given value $z^{(\tau)}$, a value of u is chosen uniformly in the region $0 \leqslant u \leqslant \tilde{p}(z^{(\tau)})$, which then defines a 'slice' through the distribution, shown by the solid horizontal lines. (b) Because it is infeasible to sample directly from a slice, a new sample of z is drawn from a region $z_{\min} \leqslant z \leqslant z_{\max}$, which contains the previous value $z^{(\tau)}$.

LATENT DIRICHLET ALLOCATION

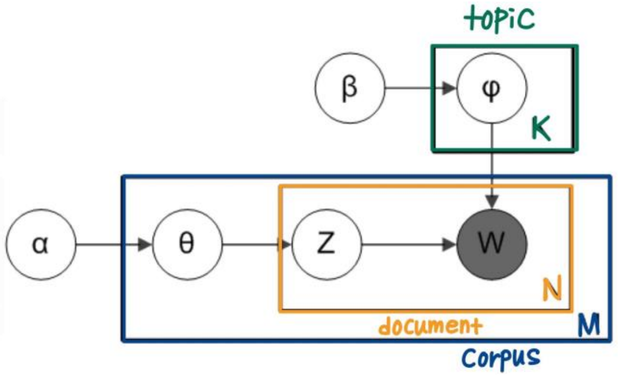
LDA – Gibbs Sampling

Review

- Graphical model
  - M: # of documents - corpus
  - N: # of words in a given document(document i has N_i words)
  - α: parameter of the Dirichlet prior on the per-document topic distributions(θ)
  - β: parameter of the Dirichlet prior on the per-topic word distribution(φ)
  - θ_i: topic distribution for document i
  - φ_k: word distribution for topic k
  - z_{ij}: topic for the j-th word in document i
  - w_{ij}: specific word



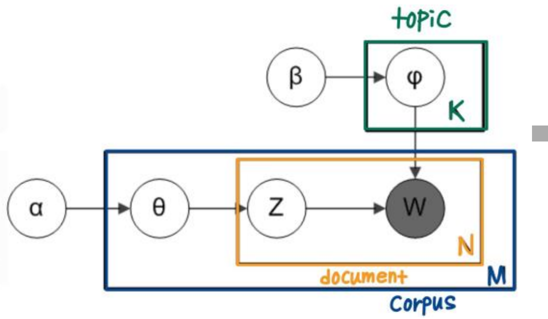
LDA – Gibbs Sampling



Review

- Generative process
  - $\theta_i \sim \mathrm{Dir}(\alpha), \quad i \in \{1, \cdots, M\}$
  - $\varphi_k \sim \mathrm{Dir}(\beta), \; k \in \{1, \cdots, K\}$
  - $z_{ij} \sim \mathrm{Multinomial}(\theta_i), \quad i \in \{1, \cdots, M\}, \quad j \in \{1, \cdots, N\}$
  - $w_{ij} \sim \mathrm{Multinomial}\left(\varphi_{z_{ij}}\right), \quad i \in \{1, \cdots, M\}, \; j \in \{1, \cdots, N\}$
  - A word w is generated from the distribution of $\varphi_{\mathbf{z}}$ word-topic distribution.
  - $\boldsymbol{z}$ topic is generated from the distribution of θ document-topic distribution.
  - θ document topic distribution is generated from the distribution of α.
  - φ word-topic distribution is generated from the distribution of β.
- We want to find the most likely allocation of $\mathbf{Z}$.
  - If we have $\boldsymbol{Z}$ distribution, we can find the most likely θ and φ.

LDA – Gibbs Sampling



Collapsed Gibbs Sampling

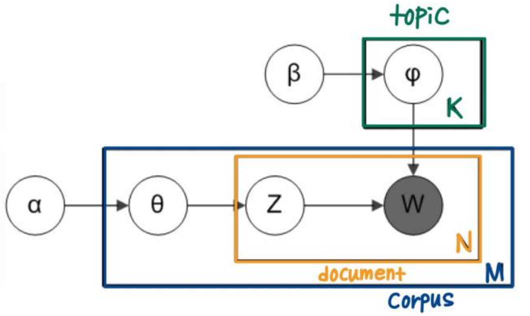
Find the most likely assignment on Z.

- Start with the factorization. $P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)=\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)P(W_{j,t}\mid\varphi_{Z_{j,t}})$,
- <u>Collapse</u> θ and φ.
  - Leave only $\mathbf{W}$(data point), $\mathbf{Z}$(sampling target), α,β(priors)
  - Marginalize out!

$$P(\boldsymbol{Z},\boldsymbol{W};\alpha,\beta)\overset{\text{marginalize}}{=}\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\varphi}}P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)\,d\boldsymbol{\varphi}\,d\boldsymbol{\theta}$$

$$\underset{\substack{\varphi\text{ and }\theta\\ \text{are independent}}}{=}\int_{\boldsymbol{\varphi}}\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{j=1}^{M}\prod_{t=1}^{N}P(W_{j,t}\mid\varphi_{Z_{j,t}})\,d\boldsymbol{\varphi}\int_{\boldsymbol{\theta}}\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)\,d\boldsymbol{\theta}.$$

LDA – Gibbs Sampling



Collapsed Gibbs Sampling

Find the most likely assignment on Z.

- Collapse θ and φ.

$$\boxed{\int_{\varphi} \prod_{i=1}^{K} P(\varphi_i;\beta) \prod_{j=1}^{M}\prod_{t=1}^{N} P(W_{j,t} \mid \varphi_{Z_{j,t}})\, d\varphi \int_{\theta} \prod_{j=1}^{M} P(\theta_j;\alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, d\boldsymbol{\theta}.}$$

  - θ part

$$\int_{\theta} \prod_{j=1}^{M} P(\theta_j;\alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, d\boldsymbol{\theta} = \prod_{j=1}^{M} \int_{\theta_j} \left\{ P(\theta_j;\alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) \right\} d\theta_j.$$

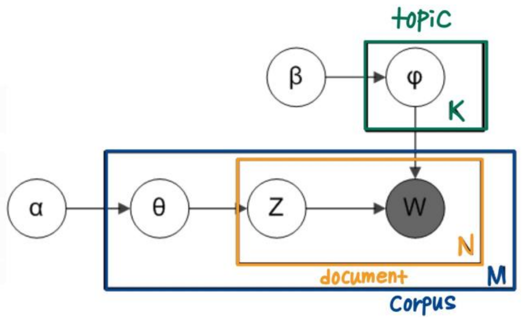
$$\int_{\theta_j} P(\theta_j;\alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, d\theta_j \underset{\theta\sim Dir(\alpha)}{=} \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_{j,i}^{\alpha_i-1} \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\, d\theta_j.$$

- $n_{j,r}^{i}$: the number of word tokens in the jth document with the same word symbol (the rth word in the (total) vocabulary) assigned to the ith topic

$$P(Z_{j,t} = i \mid \theta_j) = \theta_{j,i} \rightarrow \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) = \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i}}.$$

$$\int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_{j,i}^{\alpha_i-1} \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i}}\, d\theta_j = \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_{j,i}^{n_{j,(\cdot)}^{i}+\alpha_i-1}\, d\theta_j.$$

LDA – Gibbs Sampling



Collapsed Gibbs Sampling

Find the most likely assignment on Z.

- Collapse θ and φ.
  - θ part

$$\int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{j,i}^{\alpha_i - 1} \prod_{i=1}^K \theta_{j,i}^{n_{j,(\cdot)}^i} \, d\theta_j = \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{j,i}^{n_{j,(\cdot)}^i + \alpha_i - 1} \, d\theta_j .$$

- Use pdf property:

$$\int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i\right)}{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i)} \prod_{i=1}^K \theta_{j,i}^{n_{j,(\cdot)}^i + \alpha_i - 1} \, d\theta_j = 1.$$

pdf 적분하면 1이다.

$$p(P = \{p_i\} \mid \alpha_i) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \prod_i p_i^{\alpha_i - 1}$$

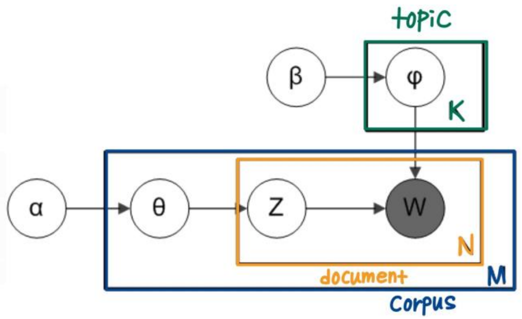
- Make Dirichlet Distribution form:

$$\int_{\theta_j} P(\theta_j; \alpha) \prod_{t=1}^N P(Z_{j,t} \mid \theta_j) \, d\theta_j = \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{j,i}^{n_{j,(\cdot)}^i + \alpha_i - 1} \, d\theta_j$$

$$= \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i)}{\Gamma\left(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i\right)} \int_{\theta_j} \underbrace{\frac{\Gamma\left(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i\right)}{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i)} \prod_{i=1}^K \theta_{j,i}^{n_{j,(\cdot)}^i + \alpha_i - 1}}_{\sim Dir(n_{j,(\cdot)}^i + \alpha_i)} \, d\theta_j$$

$$= \boxed{\frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i)}{\Gamma\left(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i\right)}}.$$

LDA – Gibbs Sampling



Collapsed Gibbs Sampling

Find the most likely assignment on Z.

- Collapse θ and φ. $$\int_{\boldsymbol{\varphi}} \prod_{i=1}^{K} P(\varphi_i;\beta) \prod_{j=1}^{M}\prod_{t=1}^{N} P(W_{j,t}\mid \varphi_{Z_{j,t}})\,d\boldsymbol{\varphi} \int_{\boldsymbol{\theta}}\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)\,d\boldsymbol{\theta}.$$
  - φ part

$$\int_{\boldsymbol{\varphi}} \prod_{i=1}^{K} P(\varphi_i;\beta) \prod_{j=1}^{M}\prod_{t=1}^{N} P(W_{j,t}\mid \varphi_{Z_{j,t}})\,d\boldsymbol{\varphi}$$

$$= \prod_{i=1}^{K}\int_{\varphi_i} P(\varphi_i;\beta) \prod_{j=1}^{M}\prod_{t=1}^{N} P(W_{j,t}\mid \varphi_{Z_{j,t}})\,d\varphi_i$$

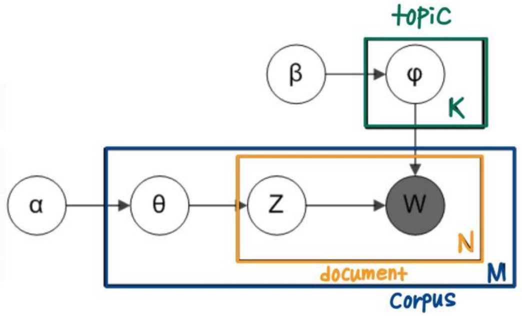
$$\underset{\varphi\sim \mathrm{Dir}(\beta)}{=} \prod_{i=1}^{K}\int_{\varphi_i} \frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)}\prod_{r=1}^{V}\varphi_{i,r}^{\beta_r-1}\prod_{r=1}^{V}\varphi_{i,r}^{n^i_{(\cdot),r}}\,d\varphi_i$$

$$= \prod_{i=1}^{K}\int_{\varphi_i} \frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)}\prod_{r=1}^{V}\varphi_{i,r}^{n^i_{(\cdot),r}+\beta_r-1}\,d\varphi_i$$

$$\underset{\text{pdf property}}{=} \boxed{\prod_{i=1}^{K}\frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)}\frac{\prod_{r=1}^{V}\Gamma(n^i_{(\cdot),r}+\beta_r)}{\Gamma\left(\sum_{r=1}^{V}n^i_{(\cdot),r}+\beta_r\right)}.}$$

$$\boxed{P(\boldsymbol{Z},\boldsymbol{W};\alpha,\beta)=\prod_{j=1}^{M}\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\frac{\prod_{i=1}^{K}\Gamma(n^i_{j,(\cdot)}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n^i_{j,(\cdot)}+\alpha_i\right)}\times\prod_{i=1}^{K}\frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)}\frac{\prod_{r=1}^{V}\Gamma(n^i_{(\cdot),r}+\beta_r)}{\Gamma\left(\sum_{r=1}^{V}n^i_{(\cdot),r}+\beta_r\right)}.}$$

LDA – Gibbs Sampling



Collapsed Gibbs Sampling

Find the most likely assignment on Z.

- Gibbs sampling formula

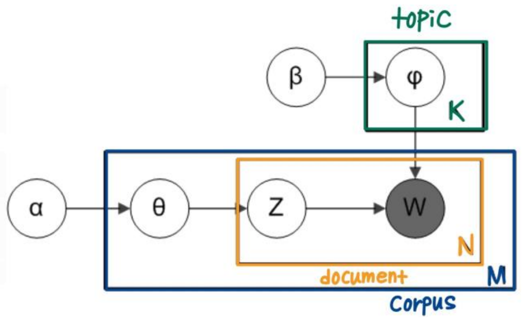
$$P(\boldsymbol{Z},\boldsymbol{W};\alpha,\beta)=\prod_{j=1}^{M}\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\frac{\prod_{i=1}^{K}\Gamma(n_{j,(\cdot)}^{i}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n_{j,(\cdot)}^{i}+\alpha_i\right)}\times\prod_{i=1}^{K}\frac{\Gamma\left(\sum_{r=1}^{V}\beta_r\right)}{\prod_{r=1}^{V}\Gamma(\beta_r)}\frac{\prod_{r=1}^{V}\Gamma(n_{(\cdot),r}^{i}+\beta_r)}{\Gamma\left(\sum_{r=1}^{V}n_{(\cdot),r}^{i}+\beta_r\right)}.$$

  - W, α, and β are assumed or data points, and $\mathbf{Z}$ is the target of sampling.
    - In Gibbs Sampling, we sample $\mathbf{Z}$ one by one.
    - The key point is to derive the following conditional probability.

$$P(Z_{(m,n)}\mid \boldsymbol{Z}_{-(m,n)},\boldsymbol{W};\alpha,\beta)=\frac{P(Z_{(m,n)},\boldsymbol{Z}_{-(m,n)},\boldsymbol{W};\alpha,\beta)}{P(\boldsymbol{Z}_{-(m,n)},\boldsymbol{W};\alpha,\beta)}$$

      - $Z_{(m,n)}$: Z hidden variable of the n-th word token in the m-th document
      - The denominator term does not affect the likelihood: $\propto P(Z_{(m,n)}=v,\boldsymbol{Z}_{-(m,n)},\boldsymbol{W};\alpha,\beta)$

LDA – Gibbs Sampling



Collapsed Gibbs Sampling

Find the most likely assignment on Z.

- Gibbs sampling formula

$$P(\boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta) = \prod_{j=1}^{M} \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \frac{\prod_{i=1}^{K} \Gamma(n_{j,(\cdot)}^{i} + \alpha_i)}{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)} \times \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \frac{\prod_{r=1}^{V} \Gamma(n_{(\cdot),r}^{i} + \beta_r)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)}.$$

$$P(Z_{(m,n)} = v \mid \boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta)$$

$$\propto P(Z_{(m,n)} = v, \boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta)$$

$$= \left(\frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)}\right)^{M} \prod_{j \neq m} \frac{\prod_{i=1}^{K} \Gamma\left(n_{j,(\cdot)}^{i} + \alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)} \left(\frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)}\right)^{K} \prod_{i=1}^{K} \left\{ \frac{\prod_{r \neq v} \Gamma\left(n_{(\cdot),r}^{i} + \beta_r\right)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)} \right\} \underbrace{\frac{\prod_{i=1}^{K} \Gamma\left(n_{m,(\cdot)}^{i} + \alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K} n_{m,(\cdot)}^{i} + \alpha_i\right)}}_{j = m} \underbrace{\prod_{i=1}^{K} \frac{\Gamma\left(n_{(\cdot),v}^{i} + \beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)}}_{r = v}$$

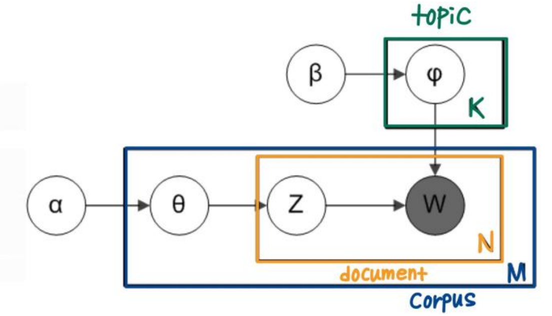
$$\propto \frac{\prod_{i=1}^{K} \Gamma\left(n_{m,(\cdot)}^{i} + \alpha_i\right)}{\Gamma\left(\sum_{i=1}^{K} n_{m,(\cdot)}^{i} + \alpha_i\right)} \prod_{i=1}^{K} \frac{\Gamma\left(n_{(\cdot),v}^{i} + \beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)}$$

j = m
r = v

$$\propto \prod_{i=1}^{K} \Gamma\left(n_{m,(\cdot)}^{i} + \alpha_i\right) \prod_{i=1}^{K} \frac{\Gamma\left(n_{(\cdot),v}^{i} + \beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)}.$$

constant term

LDA – Gibbs Sampling



Collapsed Gibbs Sampling

Find the most likely assignment on Z.

- Gibbs sampling formula
  - $n_{j,r}^{i,-(m,n)}$: the number of word tokens in the jth document with the same word symbol (the rth word in the (total) vocabulary) assigned to the ith topic
    - But with $Z_{(m,n)}$ excluded
  - Use property of gamma function: $\Gamma(x+1) = x \times \Gamma(x)$

Suppose that current word w_i is allocated to $z_i = k$

$$\prod_{i=1}^{K}\Gamma\left(n^{i}_{m,(\cdot)}+\alpha_i\right)\prod_{i=1}^{K}\frac{\Gamma\left(n^{i}_{(\cdot),v}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n^{i}_{(\cdot),r}+\beta_r\right)}.$$

$$\propto \prod_{i\neq k}\Gamma\left(n^{i,-(m,n)}_{m,(\cdot)}+\alpha_i\right)\prod_{i\neq k}\frac{\Gamma\left(n^{i,-(m,n)}_{(\cdot),v}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n^{i,-(m,n)}_{(\cdot),r}+\beta_r\right)}\Gamma\left(n^{k,-(m,n)}_{m,(\cdot)}+\alpha_k+1\right)\frac{\Gamma\left(n^{k,-(m,n)}_{(\cdot),v}+\beta_v+1\right)}{\Gamma\left(\sum_{r=1}^{V} n^{k,-(m,n)}_{(\cdot),r}+\beta_r+1\right)}$$

$$=\prod_{i\neq k}\Gamma\left(n^{i,-(m,n)}_{m,(\cdot)}+\alpha_i\right)\prod_{i\neq k}\frac{\Gamma\left(n^{i,-(m,n)}_{(\cdot),v}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n^{i,-(m,n)}_{(\cdot),r}+\beta_r\right)}\Gamma\left(n^{k,-(m,n)}_{m,(\cdot)}+\alpha_k\right)\frac{\Gamma\left(n^{k,-(m,n)}_{(\cdot),v}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n^{k,-(m,n)}_{(\cdot),r}+\beta_r\right)}\left(n^{k,-(m,n)}_{m,(\cdot)}+\alpha_k\right)\frac{n^{k,-(m,n)}_{(\cdot),v}+\beta_v}{\sum_{r=1}^{V} n^{k,-(m,n)}_{(\cdot),r}+\beta_r}$$

$$=\prod_{i}\Gamma\left(n^{i,-(m,n)}_{m,(\cdot)}+\alpha_i\right)\prod_{i}\frac{\Gamma\left(n^{i,-(m,n)}_{(\cdot),v}+\beta_v\right)}{\Gamma\left(\sum_{r=1}^{V} n^{i,-(m,n)}_{(\cdot),r}+\beta_r\right)}\left(n^{k,-(m,n)}_{m,(\cdot)}+\alpha_k\right)\frac{n^{k,-(m,n)}_{(\cdot),v}+\beta_v}{\sum_{r=1}^{V} n^{k,-(m,n)}_{(\cdot),r}+\beta_r}$$

$$\propto \left(n^{k,-(m,n)}_{m,(\cdot)}+\alpha_k\right)\frac{n^{k,-(m,n)}_{(\cdot),v}+\beta_v}{\sum_{r=1}^{V} n^{k,-(m,n)}_{(\cdot),r}+\beta_r}$$