

Sampling Based Inference

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST

icmoon@kaist.ac.kr

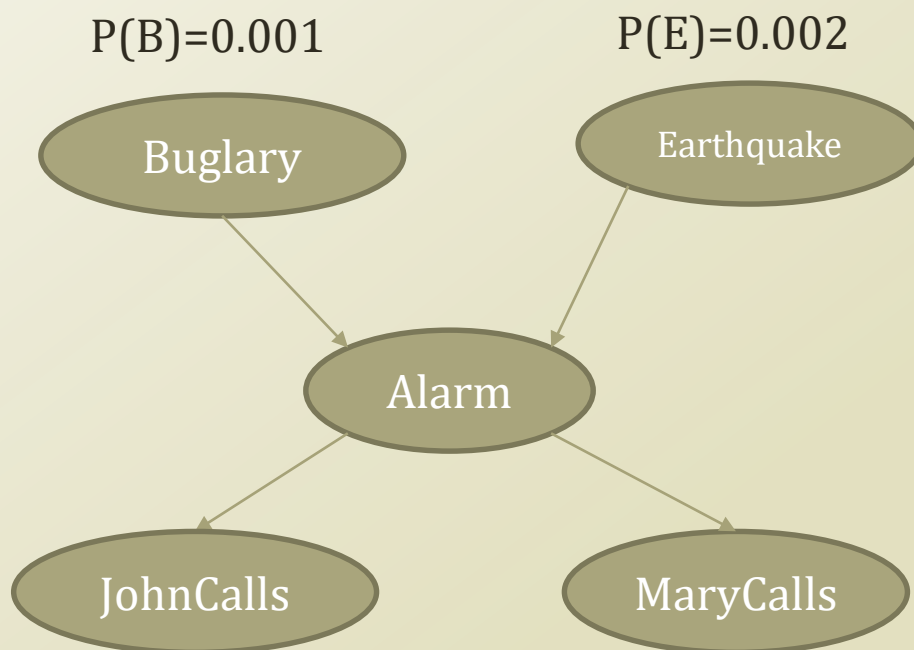
Weekly Objectives

- Learn sampling based inference method
 - Understand the concept of Markov chain Monte Carlo
 - Able to apply MCMC to the parameter inference of Bayesian networks
 - Know the mechanism of rejection sampling
 - Know the mechanism of Gibbs sampling
- Know a case study of sampling based inference
 - Understand the latent Dirichlet allocation model
 - Know the collapsed Gibbs sampling
 - Know how to derive Gibbs sampling formula for LDA

SAMPLING BASED INFERENCE

Forward Sampling

- Generate a sample from the Bayesian network
 - Follow topological order
 - Buglary \rightarrow false
 - Earthquake \rightarrow false
 - Alarm|B=F,E=F \rightarrow true
 - JC|A=T \rightarrow true
 - MC|A=T \rightarrow false
 - Create such sample many, many, many times
- Then, count the samples match the case
 - $P(E=T|MC=T)=?$
 - Count the cases of E=T and MC=T
 - Count the cases of MC=T
- Any problem?



B	E	P(A B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

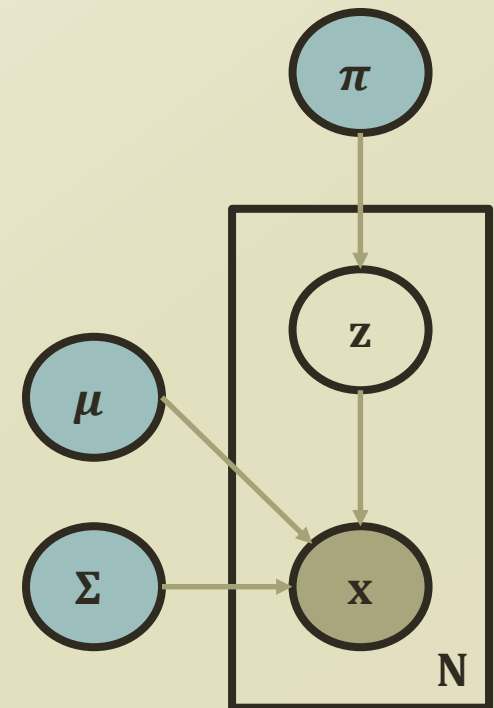
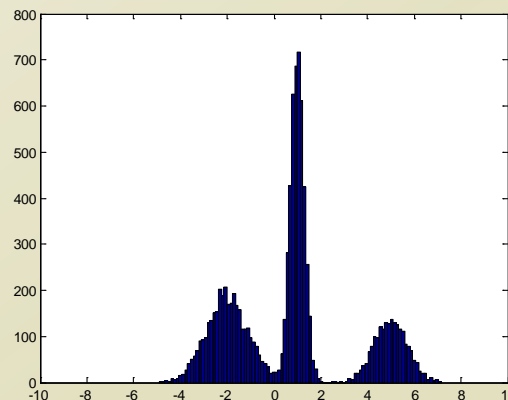
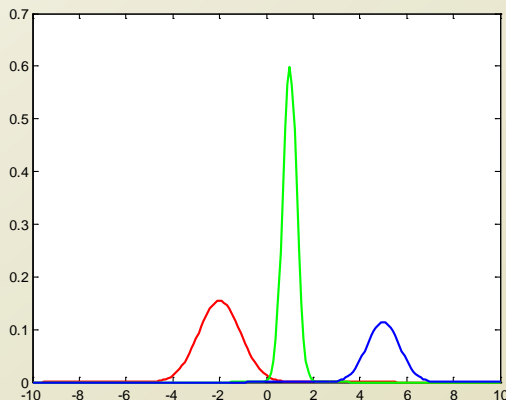
A	P(J A)
T	0.90
F	0.05

A	P(M A)
T	0.70
F	0.01

Forward Sampling in GMM

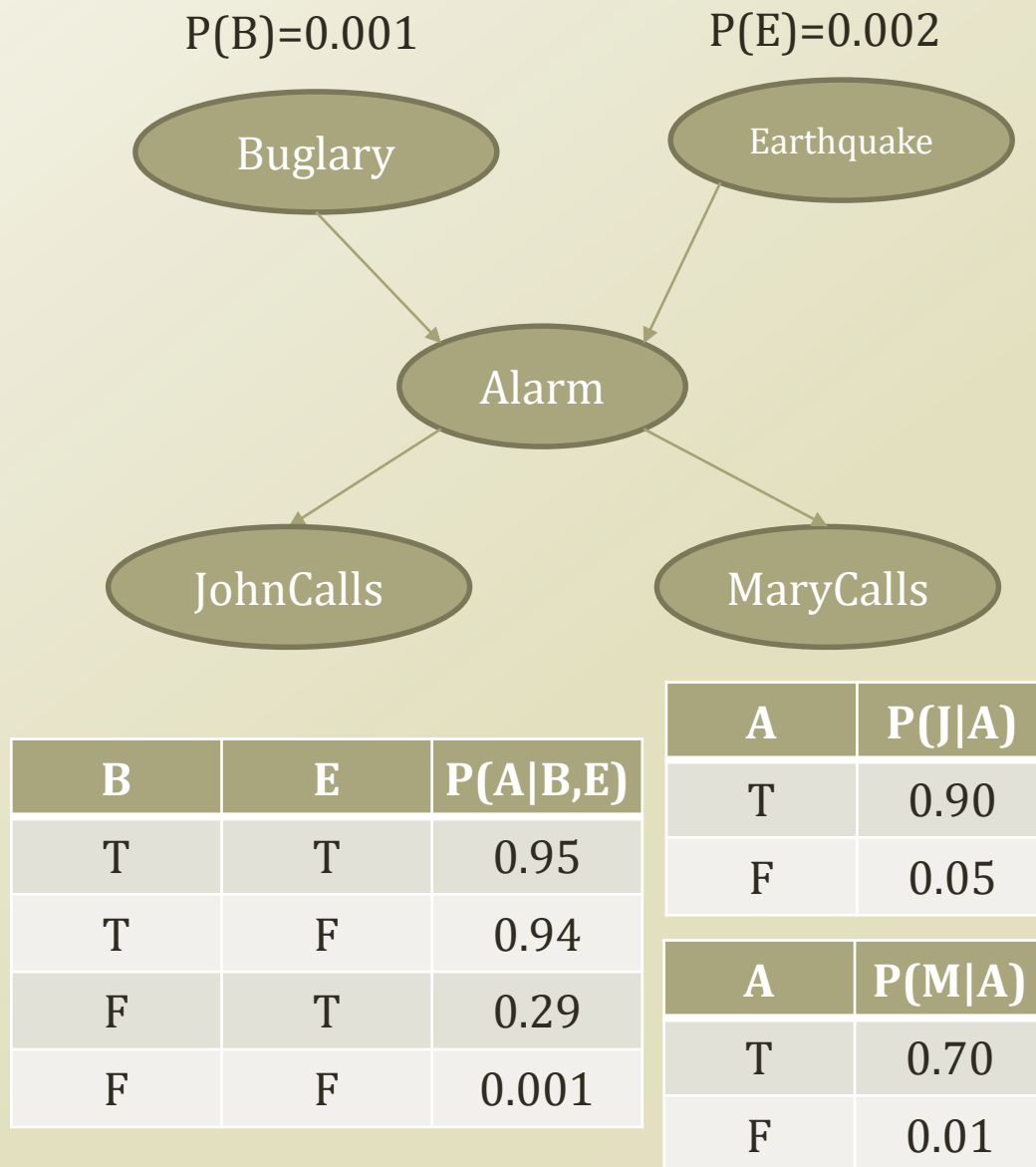
$$P(x) = \sum_{k=1}^K P(z_k)P(x|z_k) \\ = \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k)$$

- Forward sampling of GMM
 - Sample z from π
 - z is the indicator of the mixture distribution
 - With selected z , sample x from $N(\mu_z, \Sigma_z)$
- After many, many sampling, you can draw the histogram of the mixture distribution
- You have an empirical PDF, so you can ask a query like $P(0 \leq x \leq 5 | \pi, \mu, \Sigma)$



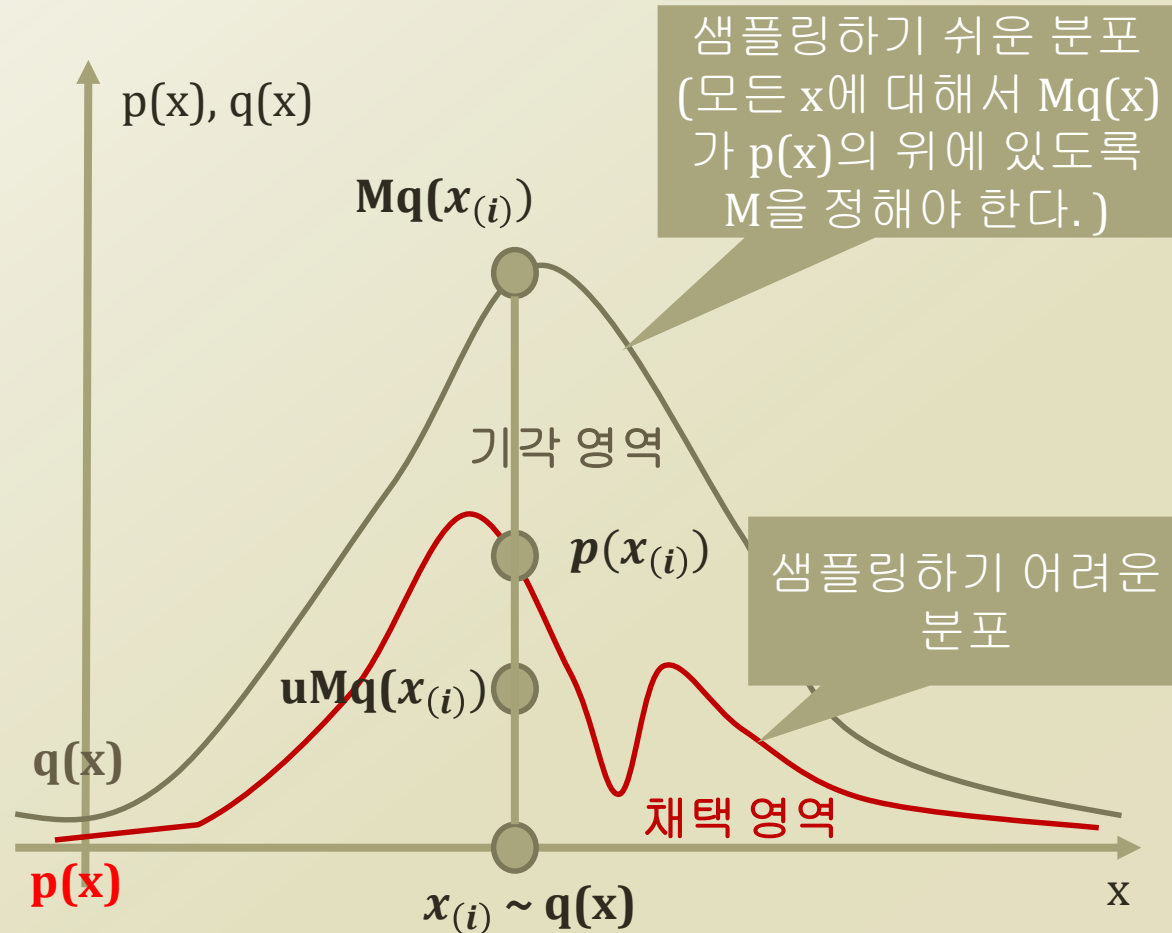
Rejection Sampling

- $P(E=T|MC=T,A=F)=?$
- RejectionSampling
 - Iterate many times
 - Generate a sample from the Bayesian network
 - Buglary \rightarrow false
 - Earthquake \rightarrow false
 - Alarm|B=F,E=F \rightarrow true
 - If the sample does not follow MC=T, A=F, reject the sampling procedure, and repeat
 - JC|A=T \rightarrow true
 - MC|A=T \rightarrow false
 - Return $\text{Count}(E=T, MC=T, A=F) / \# \text{ of Samples}$
- Any problem?



Rejection Sampling from Numerical View

- count = 0
- while count < N
 - Sample $x_{(i)} \sim q(x)$
 - Sample $u \sim \text{Unif}(0,1)$
 - If $u < \frac{p(x_{(i)})}{Mq(x_{(i)})}$
 - Accept $x_{(i)}$
 - Increase count
 - Else
 - Reject and re-sample



Rejection Sampling in GMM

$$P(x) = \sum_{k=1}^K P(z_k)P(x|z_k)$$

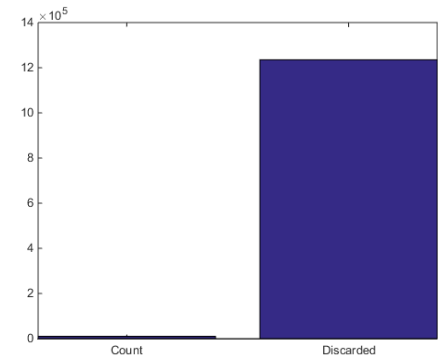
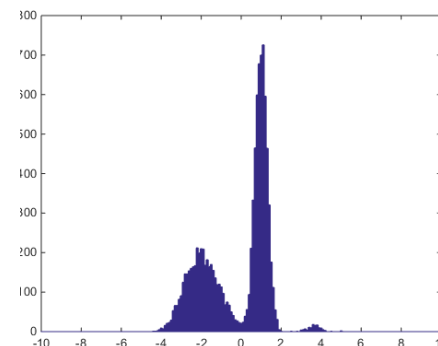
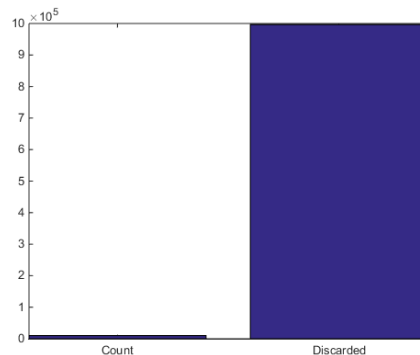
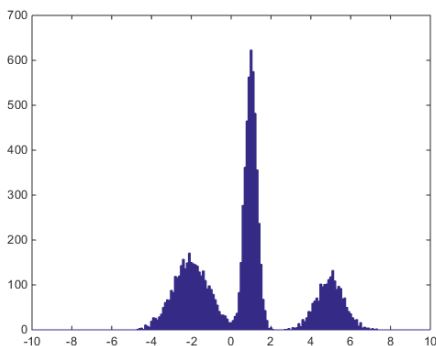
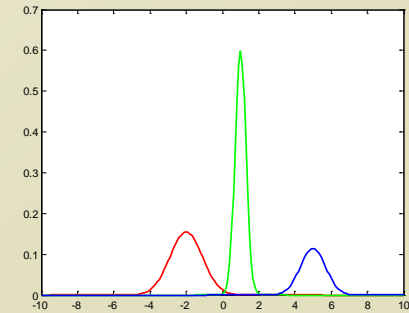
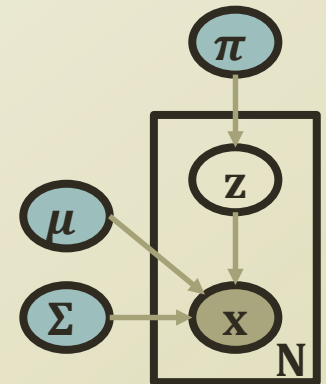
$$= \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k)$$

- Rejection sampling of GMM
 - Sample z from $\{1, 2, 3\}$ with 1/3 chance each
 - Sample x from $N(\mu_{q(z)}, \Sigma_{q(z)})$
 - $q(x)$ = The probability drawing x from $N(\mu_{q(z)}, \Sigma_{q(z)})$
 - Sample u from Uniform(0,1)
 - If $M \times u \times q(x) < p(x)$
 - Accept the sample of (z, x)
 - Else
 - Discard the sample

Q Mixture
 $= 1/3 * N(-2, 1), 1/3 * N(1, 1), 1/3 * N(5, 1)$

P Mixture
 $= 0.35 * N(-2, 0.9),$
 $0.45 * N(1, 0.3),$
 $0.2 * N(5, 0.8)$

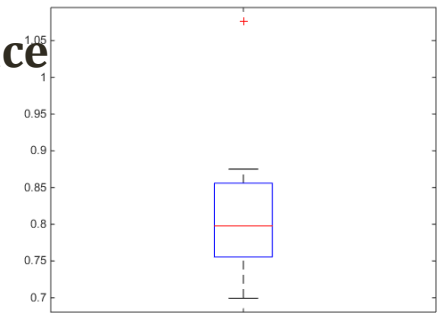
Q Mixture
 $= 3 * (1/3 * N(0, 1))$



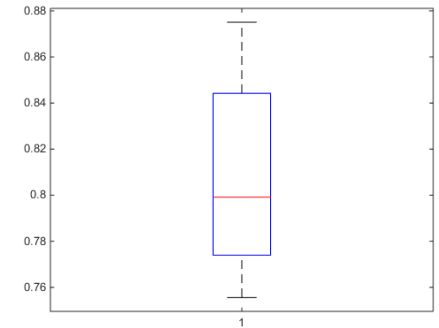
Importance Sampling

- Huge waste from the rejection
- Is generating the PDF the end goal?
 - No... Usually, the question follows
 - Calculating the expectation of PDF
 - Calculating a certain probability
- Let's use the wasted sample to answer the questions

**Importance Sampling
Prone to
Extreme
Values**



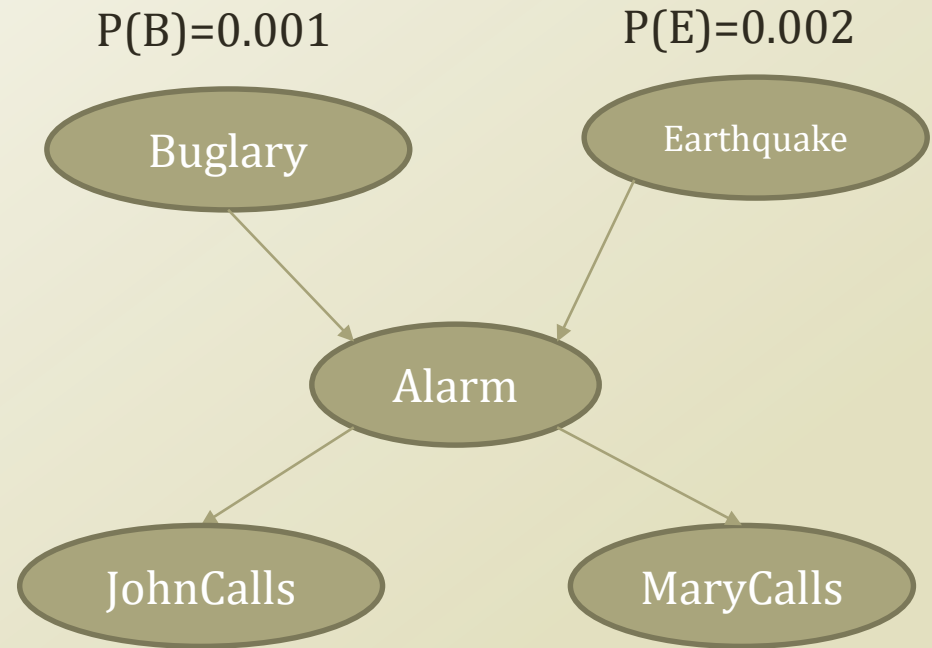
**Filtered
Extreme
Values**



- $E(f) = \int f(z)p(z)dz = \int f(z) \frac{p(z)}{q(z)} q(z)dz \cong \frac{1}{L} \sum_{l=1}^L \frac{P(z^l)}{q(z^l)} f(z^l)$
 - $L = \#$ of samples, z^l =a sample of Z
 - Here, the importance weight plays the role
 - $r^l = \frac{P(z^l)}{q(z^l)}$
 - What if $P(z^l)$ and $q(z^l)$ is not normalized, as they should be as probability distributions
 - $E(f) \cong \frac{1}{L} \sum_{l=1}^L \frac{P(z^l)}{q(z^l)} f(z^l) = \frac{1}{L} \frac{Z_q}{Z_p} \sum_{l=1}^L \frac{\tilde{P}(z^l)}{\tilde{q}(z^l)} f(z^l)$
- $P(Z>1) = \int_1^\infty 1_{z>1}p(z)dz = \int_1^\infty 1_{z>1} \frac{p(z)}{q(z)} q(z)dz \cong \frac{1}{L} \sum_{l=1}^L \frac{P(z^l)}{q(z^l)} 1_{z^l>1}$

Likelihood Weighting Algorithm

- $P(E=T|MC=T,A=F)=?$
- LikelihoodWeighting
 - SumSW=NormSW=0
 - Iterate many times
 - SW=SampleWeight = 1
 - Generate a sample from the Bayesian network
 - Buglary \rightarrow false
 - Earthquake \rightarrow false
 - Alarm=F|B=F,E=F
 - $P(A=F|B=F,E=F)=0.999$
 - $SW=1*0.999$
 - JC|A=T \rightarrow true
 - MC=T|A=F
 - $P(MC=T|A=F)=0.01$
 - $SW=1*0.999*0.01$
 - If the sample has E=T, then SumSW+=SW
 - NormSW+=SW
 - Return SumSW/NormSW
- Any further improvement?
- These samples are....



B	E	P(A B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J A)
T	0.90
F	0.05

A	P(M A)
T	0.70
F	0.01

Detour:

EM Algorithm

$$l(\theta) = \ln P(X|\theta) = \ln \left\{ \sum_Z q(Z) \frac{P(X, Z|\theta)}{q(Z)} \right\} \geq \sum_Z q(Z) \ln \frac{P(X, Z|\theta)}{q(Z)} = Q(\theta, q)$$
$$Q(\theta, q) = E_{q(Z)} \ln P(X, Z|\theta) + H(q)$$
$$L(\theta, q) = \ln P(X|\theta) - \sum_Z \left\{ q(Z) \ln \frac{q(Z)}{P(Z|X, \theta)} \right\}$$

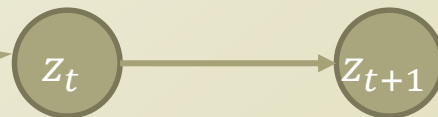
- EM algorithm
 - Finds the maximum likelihood solutions for models with latent variables
 - $P(X|\theta) = \sum_Z P(X, Z|\theta) \rightarrow \ln P(X|\theta) = \ln \{ \sum_Z P(X, Z|\theta) \}$
- EM algorithm
 - Initialize θ^0 to an arbitrary point
 - Loop until the likelihood converges
 - Expectation step
 - $q^{t+1}(z) = \operatorname{argmax}_q Q(\theta^t, q) = \operatorname{argmax}_q L(\theta^t, q) = \operatorname{argmin}_q KL(q || P(Z|X, \theta^t))$
 - $\rightarrow q^t(z) = P(Z|X, \theta) \rightarrow$ **Assign Z by $P(Z|X, \theta)$**
 - Maximization step
 - $\theta^{t+1} = \operatorname{argmax}_\theta Q(\theta, q^{t+1}) = \operatorname{argmax}_\theta L(\theta, q^{t+1})$
 - \rightarrow fixed Z means that there is no unobserved variables
 - \rightarrow Same optimization of ordinary MLE

Computing Expectation.... Sometimes, it can be hard

Detour: Markov Chain

- Markov chain
 - Each node has a probability distribution of states
 - i.e.) The probability that a state is the current state of a system
 - Concrete observation of a system: $[1 \ 0 \ 0] \rightarrow$ the system is at the first state
 - Stochastic observation of a system: $[0.7 \ 0.2 \ 0.1] \rightarrow$ the system is likely at the first state
 - The node has a vector of state probability distribution
 - Each link suggests a probabilistic state transition
 - If a system is at the first state, the probability distribution of the next state is $[0.3 \ 0.4 \ 0.3]$
 - The link has a matrix of state transition probability distribution.

$$z_t = [0.5 \quad 0.2 \quad 0.3]$$



$$\begin{aligned} P(z_{t+1}) &= P(z_t)P(z_{t+1}|z_t) = z_t T_{i,j} \\ &= [0.5 \quad 0.2 \quad 0.3] \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.4 & 0.2 & 0.4 \end{bmatrix} \\ &= [0.51 \quad 0.22 \quad 0.27] \end{aligned}$$

- The system has three states, a, b, and c.
- Transition matrix is

$$P(z_j|z_i) = T_{i,j} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}$$

Detour: Properties of Markov Chain

- Accessible
 - $i \rightarrow j$: State j is **accessible** from i if $T_{i,j}^k > 0$ and $k \geq 0$
 - $i \leftrightarrow j$: State i and j **communicate** if $i \rightarrow j$ and $j \rightarrow i$
- Reducibility
 - A Markov chain is **irreducible** if $i \leftrightarrow j, \forall i \in S, \forall j \in S$
- Periodicity
 - State i has **period** d if $d = \gcd\{n: T_{i,i}^n > 0\}$
 - If $d=1$, State i is **aperiodic**.
- Transience
 - State j is **recurrent** if $P(\inf(t \geq 1: X_t = j) < \infty | X_0 = j) = 1$
 - States which are not **recurrent** are **transient**.
- Ergodicity
 - A state is **ergodic** if the state is (positive) **recurrent** and **aperiodic**.
 - Markov chain is ergodic if all states are ergodic.

Detour: Stationary Distribution

- $RT_i = \min\{n > 0: X_n = i | X_0 = i\}$
 - Return time to state i after the departure from state i
- Limit theorem of Markov chain
 - A friend in ISE dept. told me.....
 - If a Markov chain is irreducible and ergodic
 - $\pi_i = \lim_{n \rightarrow \infty} T_{i,j}^{(n)} = \frac{1}{E[RT_i]}$
 - π_i is uniquely determined by the set of equations
 - $\pi_i \geq 0, \sum_{i \in S} \pi_i = 1, \pi_j = \sum_{i \in S} \pi_i T_{i,j}$
 - How to compute π given T
 - $\pi(I_{|S|,|S|} - T + 1_{|S|,|S|}) = 1_{1,|S|}$
 - $\pi_j = \sum_{i \in S} \pi_i T_{i,j} \rightarrow \pi_j - \sum_{i \in S} \pi_i T_{i,j} = 0 \rightarrow \pi(I_{|S|,|S|} - T) = 0$
 - To the above formula, apply $\sum_{i \in S} \pi_i = 1 \rightarrow \pi 1_{|S|,|S|} = 1_{1,|S|}$ to both sides
 - $\pi(I_{|S|,|S|} - T + 1_{|S|,|S|}) = 1_{1,|S|}$
 - Here, π is the stationary distribution!

```
>> T
T =
    0.7000    0.2000    0.1000
    0.2000    0.3000    0.5000
    0.4000    0.2000    0.4000

>> pi = ones(1,3) / (eye(3,3)-T+ones(3,3))
pi =
    0.5079    0.2222    0.2698

>> pi*T
ans =
    0.5079    0.2222    0.2698

>> pi(1)*T(1,2)
ans =
    0.1016

>> pi(2)*T(2,1)
ans =
    0.0444
```

Reversible MC

```
>> T2 = [0 0.5 0.5 ; 0.25 0.5 0.25; 0.25 0.
T2 =
         0    0.5000    0.5000
    0.2500    0.5000    0.2500
    0.2500    0.2500    0.5000

>> pi2 = ones(1,3)/(eye(3,3)-T2+ones(3,3))
pi2 =
    0.2000    0.4000    0.4000

>> pi2*T2
ans =
    0.2000    0.4000    0.4000

>> pi2(1)*T2(1,2)
ans =
    0.1000

>> pi2(2)*T2(2,1)
ans =
    0.1000
```

Irreversible MC

π is the stationary distribution



Reversible Markov chain

$$\pi_i T_{i,j} = \pi_j T_{j,i}$$

Detailed Balance, or
Balance Equation

Markov Chain for Sampling

- Problem of the previous samplings?
 - No use of the past records → every sampling is independent
- Assigning Z values is a key in the inference
 - Let's assign the values by sampling result
 - Calculate $P(E|MC=T,A=F)$ → Toss a biased coin to assign a value to E
- Sequence of random variables such a process moves through, with the Markov property defining serial dependence only between adjacent periods (as in a "chain")
- A Markov chain is a stochastic process with the Markov property
 - Example) First-order Markov chain



- $p(z^{(m+1)} | z^{(1)}, \dots, z^{(m)}) = p(z^{(m+1)} | z^{(m)}), m \in \{1, \dots, M - 1\}$
- Describing systems that follow a chain of linked events, where what happens next depends only on the current state of the system

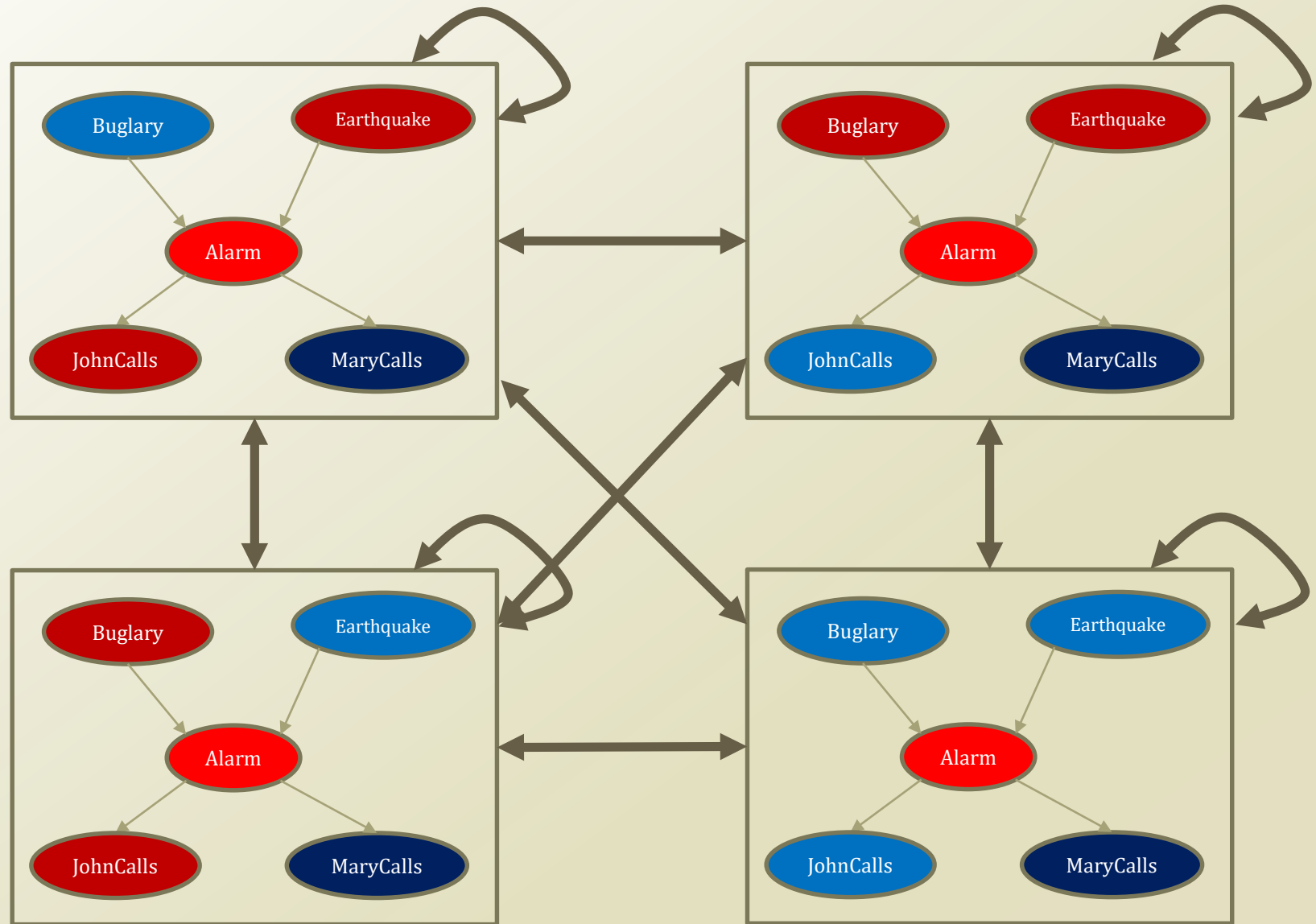
Markov chain theory vs. Markov Chain Monte Carlo



- Traditional Markov Chain analysis :
 - A transition rule, $p(z^{(t+1)} | z^{(t)})$, is given,
 - Interested in finding the stationary distribution $\pi(z)$

- Markov chain Monte Carlo(MCMC) :
 - A target stationary distribution $\pi(z)$ is known,
 - Interested in prescribing an efficient transition rule to reach the stationary distribution
 - Algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution $\pi(z)$
 - Starting from an arbitrary state, the Markov chain proceeds
 - $\underbrace{z^{(1)} \rightarrow z^{(2)} \rightarrow \dots \rightarrow z^{(m)}}_{\text{Burn-in period}} \rightarrow \underbrace{z^{(m+1)} \rightarrow z^{(m+2)} \rightarrow \dots \rightarrow z^{(m+n)}}_{\text{Treat them as samples from } \pi(x)}$

Markov Chain of Z



$q(z^t|z^*)P(z^*) < q(z^*|z^t)P(z^t) \rightarrow$ Movement from z^t to z^* is too often

Metropolis-Hastings Algorithm

- General algorithm of MCMC

- Current value: z^t
- Propose a candidate $z^* \sim q(z^*|z^t)$ where q_t is a proposal distribution
 - Same as importance and rejection samplings, yet the difference is the Markov property idea in the proposal distribution
- With an acceptance probability, α
 - Accept $\rightarrow z^{t+1} = z^*$
 - Reject $\rightarrow z^{t+1} = z^t$

- Metropolis-Hastings algorithm

- Given the general algorithm of MCMC
- Consider a ratio, $r(z^*|z^t) = \frac{q(z^t|z^*)P(z^*)}{q(z^*|z^t)P(z^t)}$, we want this to be 1
 - $q(z^t|z^*)P(z^*)r_{z^* \rightarrow z^t} = q(z^*|z^t)P(z^t)r_{z^t \rightarrow z^*}$
 - $r(z^*|z^t) < 1 \rightarrow q(z^t|z^*)P(z^*) < q(z^*|z^t)P(z^t)$
 - Increase $r_{z^* \rightarrow z^t} = 1$, decrease $r_{z^t \rightarrow z^*} = r(z^*|z^t)$
 - $r(z^*|z^t) > 1 \rightarrow q(z^t|z^*)P(z^*) > q(z^*|z^t)P(z^t)$
 - Decrease $r_{z^* \rightarrow z^t} = r(z^t|z^*)$, increase $r_{z^t \rightarrow z^*} = 1$
- Acceptance probability $\alpha(z^*|z^t) = \min\{1, r(z^*|z^t)\}$

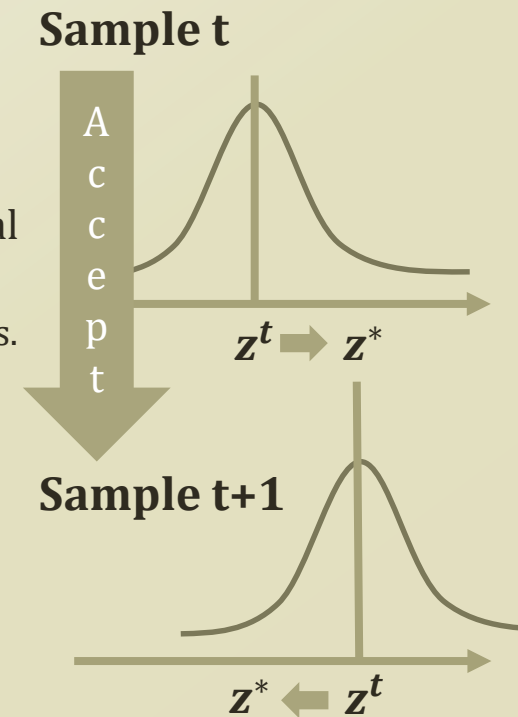
We want the stationary distribution, $\pi(z)$, of our MCMC sampling to be $P(z)$

Reversible Markov chain
 $\pi_i T_{i,j} = \pi_j T_{j,i}$

q is not well-designed to be the reversible MC, so we adjust by r

Random Walk M-H Algorithm

- $T_{t,*}^{MH} = q(z^*|z^t)\alpha(z^*|z^t)$
 - Transition probability to satisfy the balance equation with $P(z)$ as the stationary distribution
 - $\alpha(z^*|z^t) = \min\{1, \frac{q(z^t|z^*)P(z^*)}{q(z^*|z^t)P(z^t)}\}$
 - Here, we already assumed, so far, the easy calculation of $P(z)$
 - What we miss is the definition of $q(z^*|z^t)$, but this is a proposal that any probability distribution can be
 - Surely, there are better and worse proposal probability distributions.
 - Choosing $q(z^*|z^t)$ determines the type of M-H algorithm
- Random walk M-H algorithm
 - $T_{t,*}^{MH} = q(z^*|z^t)\alpha(z^*|z^t)$
 - $z^* \sim N(z^t, \sigma^2)$
 - $q(z^*|z^t) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(z^*-z^t)^2}{2\sigma^2})$



Random Walk Process

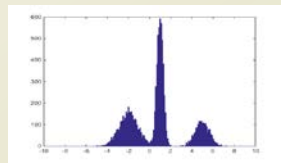
Result of Random Walk M-H

Sampling Result of Random Walk M-H

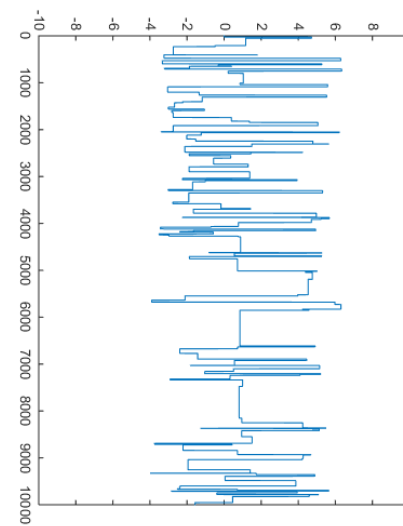
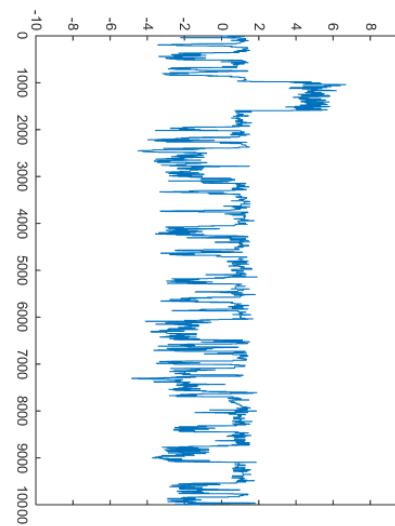
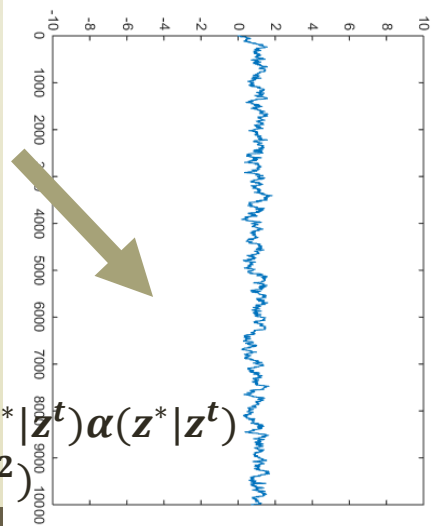
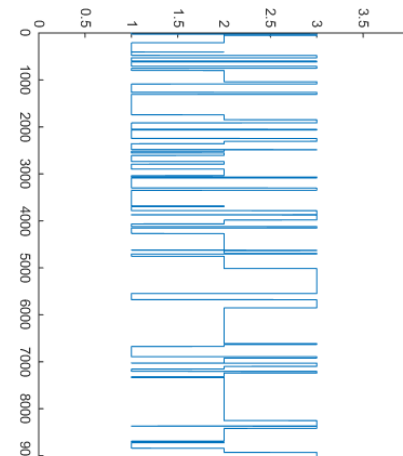
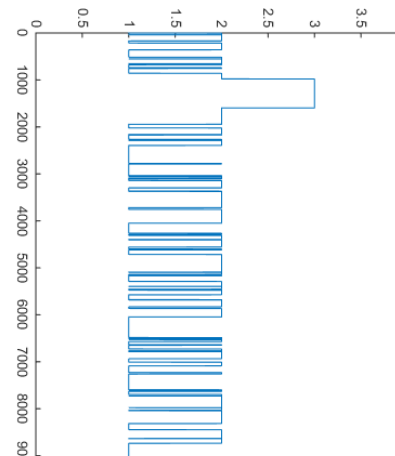
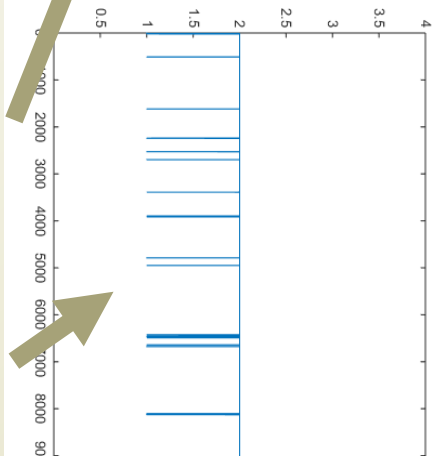
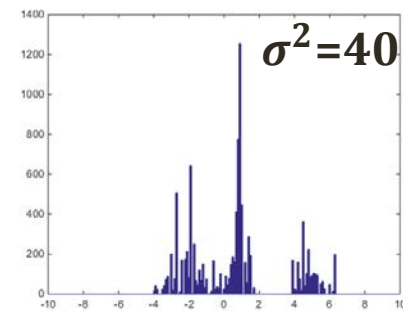
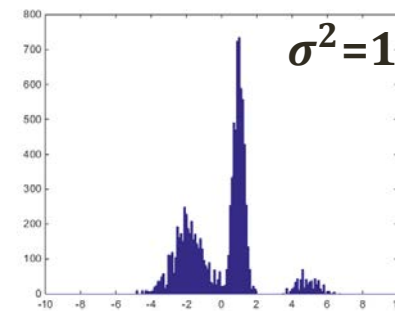
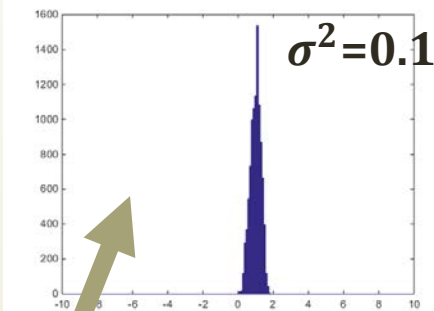
Overall Sampling

Latent Mode Selection Sampling

Observed Variable Sampling

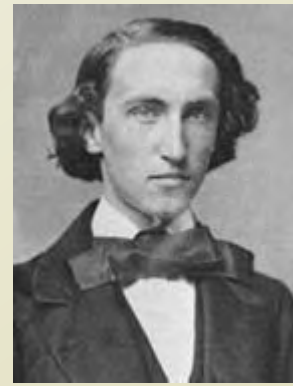


Target Mixture Distribution $T_{t,*}^{MH} = q(z^* | z^t) \alpha(z^* | z^t)$
 $z^* \sim N(z^t, \sigma^2)$

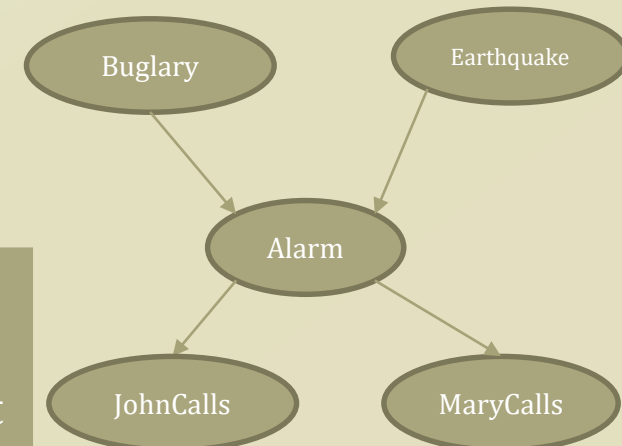
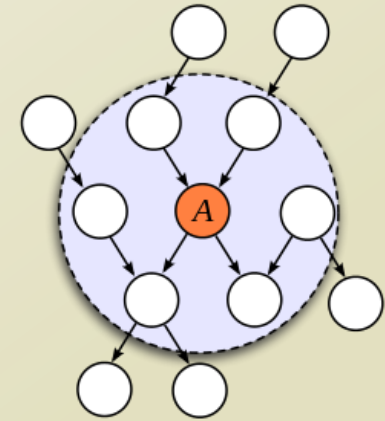


Gibbs Sampling

Josiah Willard Gibbs
(1839 - 1903)
- physicist



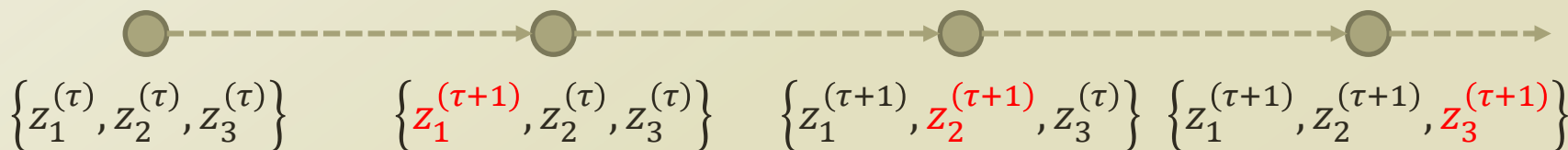
- Gibbs Sampling: A special case of M-H algorithm
 - Let's suppose $z^t = (z_k^t, z_{-k}^t) \rightarrow z^* = (z_k^*, z_{-k}^t)$
 - $T_{t,*}^{MH} = q(z^*|z^t)\alpha(z^*|z^t)$
 - $q(z^*|z^t) = P(z_k^*, z_{-k}^t|z_{-k}^t) = P(z_k^*|z_{-k}^t)$
 - Let's observe the balance equation
 - Should hold $P(z^t)q(z^*|z^t) = P(z^*)q(z^t|z^*)$
 - $P(z^t)q(z^*|z^t) = P(z_k^t, z_{-k}^t)P(z_k^*|z_{-k}^t) = P(z_k^t|z_{-k}^t)P(z_{-k}^t)P(z_k^*|z_{-k}^t) = P(z_k^t|z_{-k}^t)P(z_k^*, z_{-k}^t) = q(z^t|z^*)P(z^*)$
 - Always hold the balance equation!
 - Then, the acceptance probability becomes $\alpha(z^*|z^t) = 1$
- Example of Gibbs sampling
 - When the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is easy
 - $P(E, J, C, B | A=F, MC=T)=?$
 - Hard to sample directly. Why?
 - Consider a conditional distribution $p(z_i|z_{-i}, e)$
 - $P(E|B, A, J, C, MC)=P(E|A, B)$
 - $P(J|C, B, E, A, MC)=P(J|C, A)$
 - $P(B|E, A, J, C, MC)=P(B|A, E)$
 - Update one random variable at a time



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
- Example
 1. Full joint probability : $p(z_1, z_2, z_3)$
 2. Sample $z_1 \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}) \rightarrow$ Obtain a value $z_1^{(\tau+1)}$
 3. Sample $z_2 \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}) \rightarrow$ Obtain a value $z_2^{(\tau+1)}$
 4. Sample $z_3 \sim p(z_3 | z_1^{(\tau+1)}, z_2^{(\tau+1)}) \rightarrow$ Obtain a value $z_3^{(\tau+1)}$

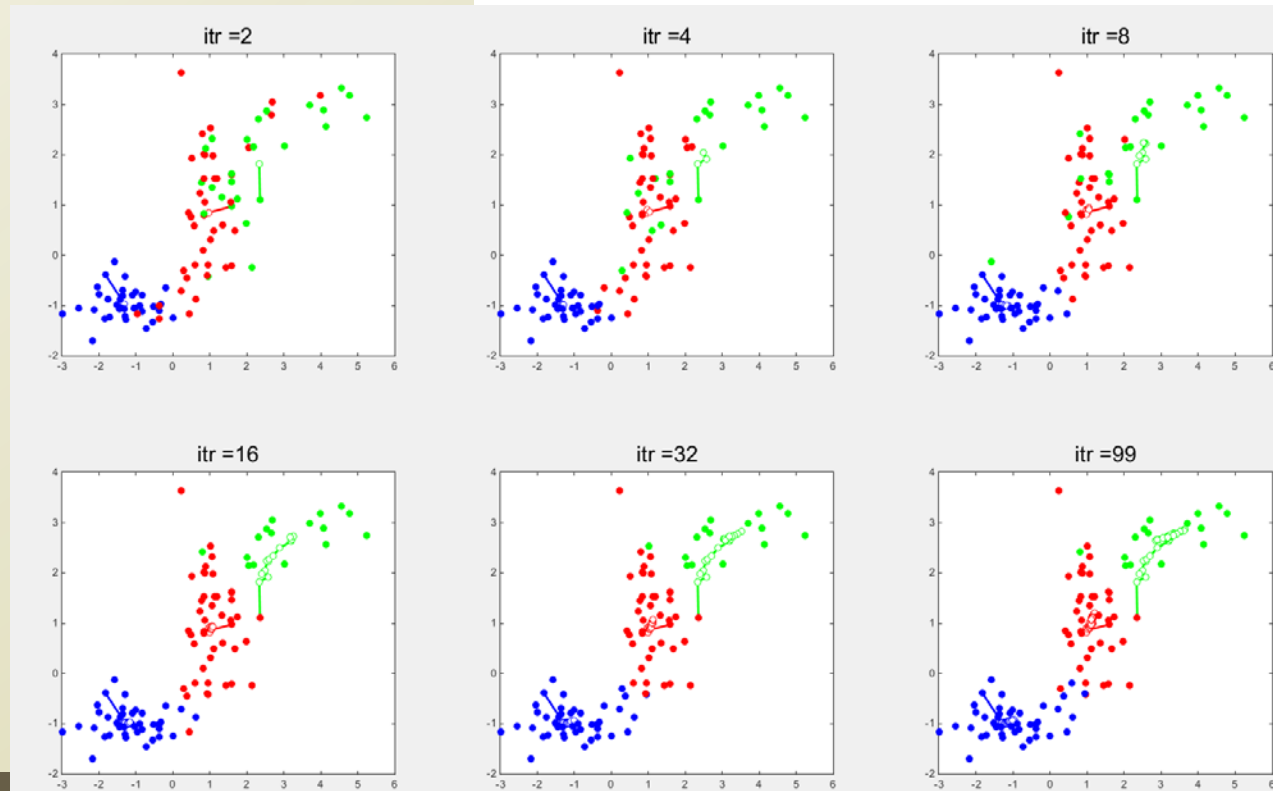
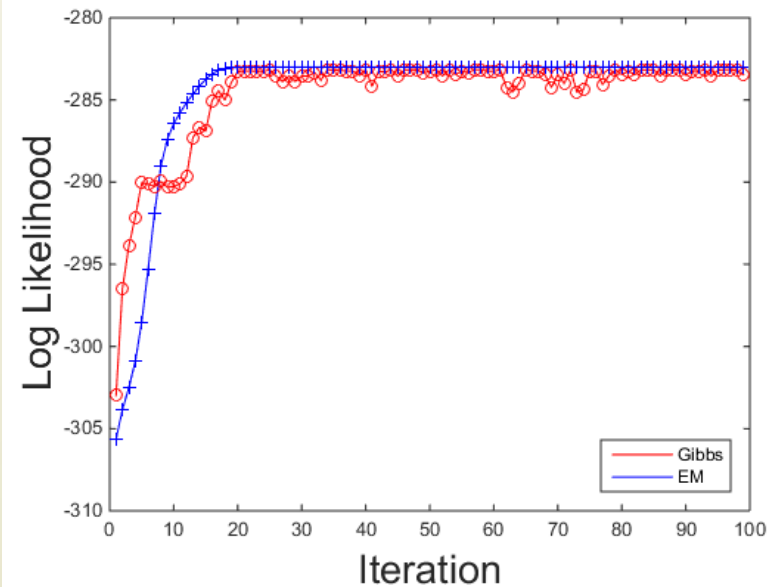


Gibbs Sampling Algorithm

- Full joint distribution, $p(\mathbf{z}) = p(z_1, \dots, z_M)$
- State = $\{z_i: i = 1, \dots, M\}$
- Algorithm
 1. Initialize $\{z_i: i = 1, \dots, M\}$
 2. For step $\tau = 1, \dots, T$:
 - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \dots, z_M^{(\tau)})$
 - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \dots, z_M^{(\tau)})$
 - \vdots
 - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \dots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \dots, z_M^{(\tau)})$
 - \vdots
 - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \dots, z_{M-1}^{(\tau+1)})$

Gibbs Sampling based GMM

- Hard to tell the performance with the simple GMM
 - Sampling based inference
 - Simulation based
 - EM based inference
 - Optimization based
- Real power of Gibbs sampler comes from collapsing! → Collapsed Gibbs Sampler
 - Let's look at more sophisticated model for collapsing technique

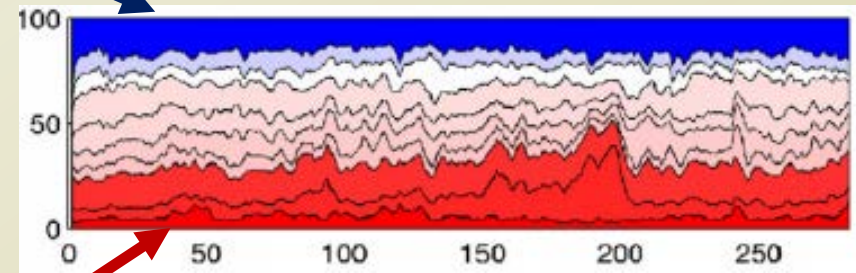


LATENT DIRICHLET ALLOCATION

Topic Modeling

- Sample results of topic modeling on News paper searched by 'Obama'

family, women, home, life, told, news, open, night
school, students, education, energy, oil, schools, program, district
tax, percent, budget, spending, cuts, economy, jobs, debt
department, agency, office, report, general, information, service
health, financial, companies, insurance, business, company, care
iran, nuclear, international, foreign, russia, syria, countries, minister
senate, congress, republicans, court, democrats, vote, committee
policy, world, rights, americans, change, speech, important, today
military, afghanistan, war, iraq, security, forces, defense, troops
romney, campaign, party, republican, voters, election, republicans

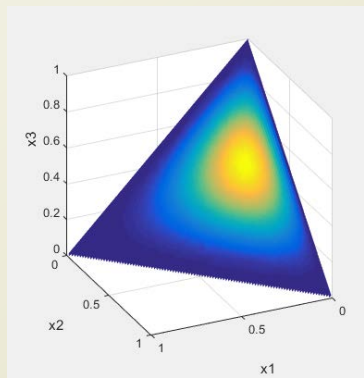


Days

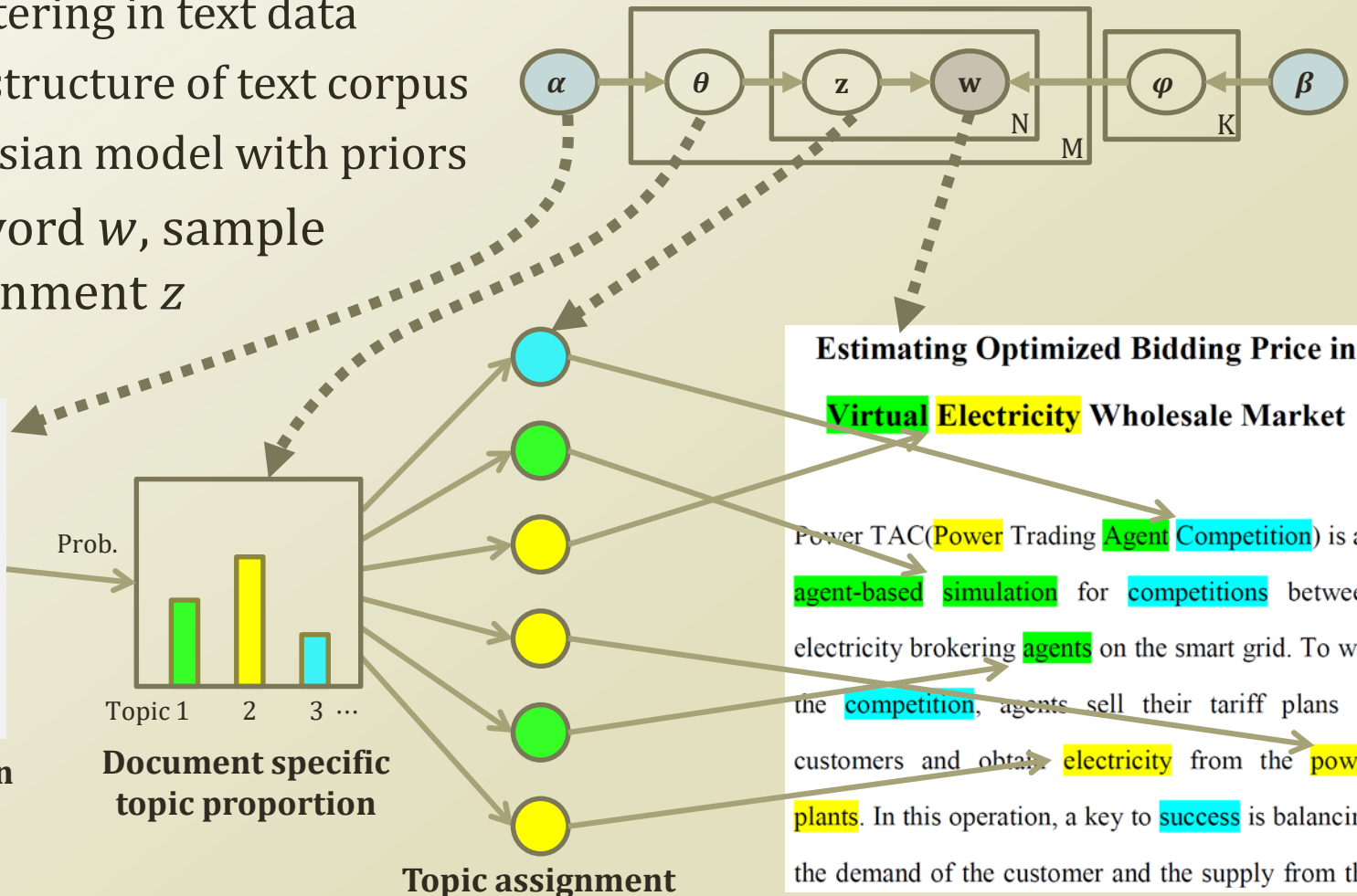
- What we can identify are
 - Topics
 - The proportion of topics
 - The most probable words in topics
- Text analysis without reading the whole corpus

Latent Dirichlet Allocation

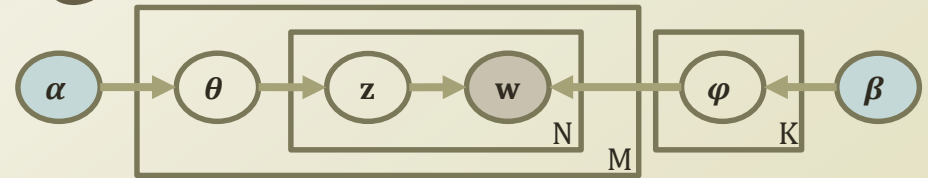
- Latent Dirichlet Allocation
 - Soft clustering in text data
 - Has the structure of text corpus
 - Is a Bayesian model with priors
- For each word w , sample topic assignment z



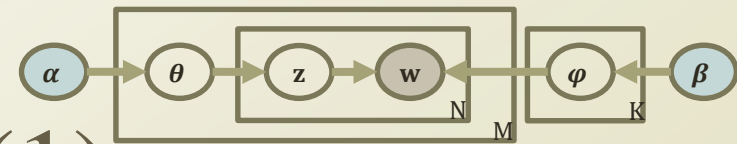
Dirichlet Distribution Prior



Finding Topic Assignment Per Word



- Let's treat this as a Bayesian network
 - Do you remember the story of "Alarm and call"?
 - There was a story of **generating** Mary's call from the event
 - Generative Process**
 - $\theta_i \sim \text{Dir}(\alpha), i \in \{1, \dots, M\}$
 - $\varphi_k \sim \text{Dir}(\beta), k \in \{1, \dots, K\}$
 - $z_{i,l} \sim \text{Mult}(\theta_i), i \in \{1, \dots, M\}, l \in \{1, \dots, N\}$
 - $w_{i,l} \sim \text{Mult}(\varphi_{z_{i,l}}), i \in \{1, \dots, M\}, l \in \{1, \dots, N\}$
 - A word w is generated from the distribution of φ_z word-topic distribution
 - 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
- If we have Z distribution, we can find the most likely θ and φ
 - θ : Topic distribution in a document
 - φ : Word distribution in a topic
 - Finding the most likely allocation of Z is the key of inference on θ and φ



Gibbs Sampling on Z (1)

- Finding the most likely assignment on Z → Gibbs Sampling
- Start with the factorization

- $P(W, Z, \theta, \varphi; \alpha, \beta)$

$$= \prod_{i=1}^K P(\varphi_i; \beta) \prod_{j=1}^M P(\theta_j; \alpha) \prod_{l=1}^N P(Z_{j,l} | \theta_j) P(W_{j,l} | \varphi_{Z_{j,l}})$$

- We are going to collapse θ and φ to leave only W, Z, α and β
 - Why? W (Data point), Z (Sampling Target), α and β (priors)
 - Collapsed Gibbs sampling!

- $P(W, Z; \alpha, \beta) = \int_{\theta} \int_{\varphi} P(W, Z, \theta, \varphi; \alpha, \beta) d\varphi d\theta =$

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

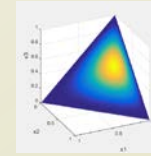
$$\int_{\theta} \prod_{j=1}^M P(\theta_j; \alpha) \prod_{l=1}^N P(Z_{j,l} | \theta_j) d\theta$$

1. Independence between two integrals
2. Need to remove the integrals and come up with the sampling distribution

Gibbs Sampling on Z (2)

$$x \sim \text{Dir}(\alpha)$$

$$P(X|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

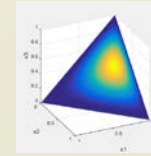


- $P(W, Z; \alpha, \beta) = \int_{\theta} \int_{\varphi} P(W, Z, \theta, \varphi; \alpha, \beta) d\varphi d\theta =$
 $\int_{\varphi} \prod_{i=1}^K P(\varphi_i; \beta) \prod_{j=1}^M \prod_{l=1}^N P(W_{j,l} | \varphi_{Z_{j,l}}) d\varphi \times \int_{\theta} \prod_{j=1}^M P(\theta_j; \alpha) \prod_{l=1}^N P(Z_{j,l} | \theta_j) d\theta = (1) \times (2)$
- $(1) = \int_{\varphi} \prod_{i=1}^K P(\varphi_i; \beta) \prod_{j=1}^M \prod_{l=1}^N P(W_{j,l} | \varphi_{Z_{j,l}}) d\varphi =$
 $\prod_{i=1}^K \int_{\varphi_i} P(\varphi_i; \beta) \prod_{j=1}^M \prod_{l=1}^N P(W_{j,l} | \varphi_{Z_{j,l}}) d\varphi_i =$
 $\prod_{i=1}^K \int_{\varphi_i} \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \prod_{v=1}^V \varphi_{i,v}^{\beta_v - 1} \prod_{j=1}^M \prod_{l=1}^N P(W_{j,l} | \varphi_{Z_{j,l}}) d\varphi_i$
 - We introduce a new count of $n_{j,r}^i$: number of words assigned to i-th topic in j-th document with r-th unique word
- $= \prod_{i=1}^K \int_{\varphi_i} \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \prod_{v=1}^V \varphi_{i,v}^{\beta_v - 1} \prod_{v=1}^V \varphi_{i,v}^{n_{(\cdot),v}^i} d\varphi_i$
- $= \prod_{i=1}^K \int_{\varphi_i} \frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \prod_{v=1}^V \varphi_{i,v}^{n_{(\cdot),v}^i + \beta_v - 1} d\varphi_i$
- $= \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v) \Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v) \Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)} \int_{\varphi_i} \frac{\Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)}{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v)} \prod_{v=1}^V \varphi_{i,v}^{n_{(\cdot),v}^i + \beta_v - 1} d\varphi_i$
- $= \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v) \Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v) \Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)}$

Gibbs Sampling on Z (3)

$$x \sim \text{Dir}(\alpha)$$

$$P(X|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i-1}$$



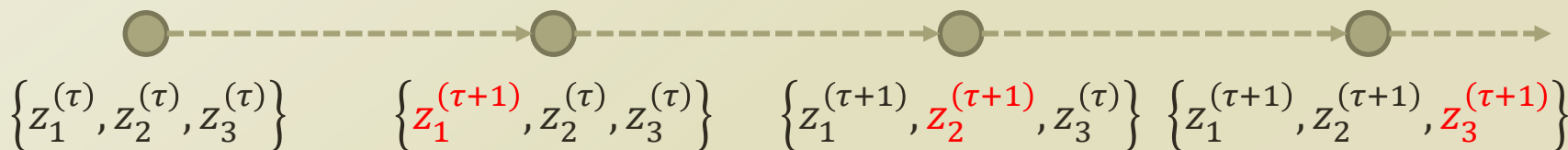
- $P(W, Z; \alpha, \beta) = \int_{\theta} \int_{\varphi} P(W, Z, \theta, \varphi; \alpha, \beta) d\varphi d\theta =$
 $\int_{\varphi} \prod_{i=1}^K P(\varphi_i; \beta) \prod_{j=1}^M \prod_{l=1}^N P(W_{j,l} | \varphi_{Z_{j,l}}) d\varphi \times \int_{\theta} \prod_{j=1}^M P(\theta_j; \alpha) \prod_{l=1}^N P(Z_{j,l} | \theta_j) d\theta = (1) \times (2)$
- $(2) = \int_{\theta} \prod_{j=1}^M P(\theta_j; \alpha) \prod_{l=1}^N P(Z_{j,l} | \theta_j) d\theta = \prod_{j=1}^M \int_{\theta_j} P(\theta_j; \alpha) \prod_{l=1}^N P(Z_{j,l} | \theta_j) d\theta_j =$
 $\prod_{j=1}^M \int_{\theta_j} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{j,k}^{\alpha_k-1} \prod_{l=1}^N P(Z_{j,l} | \theta_j) d\theta_j$
 - We introduce a new count of $n_{j,r}^i$: number of words assigned to i-th topic in j-th document with r-th unique word
- $= \prod_{j=1}^M \int_{\theta_j} \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{j,i}^{\alpha_i-1} \prod_{k=1}^K \theta_{j,k}^{n_{j,(.)}^k} d\theta_j$
- $= \prod_{j=1}^M \int_{\theta_j} \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_{j,i}^{n_{j,(.)}^i + \alpha_i - 1} d\theta_j$
- $= \prod_{j=1}^M \frac{\prod_{i=1}^K \Gamma(n_{j,(.)}^i + \alpha_i) \Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i) \Gamma(\sum_{i=1}^K n_{j,(.)}^i + \alpha_i)} \int_{\theta_j} \frac{\Gamma(\sum_{i=1}^K n_{j,(.)}^i + \alpha_i)}{\prod_{i=1}^K \Gamma(n_{j,(.)}^i + \alpha_i)} \prod_{i=1}^K \theta_{j,i}^{n_{j,(.)}^i + \alpha_i - 1} d\theta_j$
- $= \prod_{j=1}^M \frac{\prod_{i=1}^K \Gamma(n_{j,(.)}^i + \alpha_i) \Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i) \Gamma(\sum_{i=1}^K n_{j,(.)}^i + \alpha_i)}$

Collapse from Conjugacy

- Same mechanism to remove θ and φ
 - $$= \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v) \Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v) \Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)} \int_{\varphi_i} \frac{\Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)}{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v)} \prod_{v=1}^V \varphi_{i,v}^{n_{(\cdot),v}^i + \beta_v - 1} d\varphi_i$$
 - $$= \prod_{j=1}^M \frac{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_k) \Gamma(\sum_{i=1}^K \alpha_k)}{\prod_{i=1}^K \Gamma(\alpha_i) \Gamma(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_k)} \int_{\theta_j} \frac{\Gamma(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_k)}{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_k)} \prod_{i=1}^K \theta_{j,i}^{n_{j,(\cdot)}^i + \alpha_k - 1} d\theta_j$$
- This is a multiplication of the Dirichlet distribution and the multinomial distribution. After multiplication, another Dirichlet distribution emerges.
 - In LDA: i.e. $\int_{\theta} \prod_{j=1}^M P(\theta_j; \alpha) \prod_{l=1}^N P(Z_{j,l} | \theta_j) d\theta$
 - In general: $P(X|\theta) \times P(\theta)$
 - Likelihood and prior multiplication results in the prior distribution \rightarrow Conjugate prior
- LDA utilizes the multinomial distribution and the Dirichlet distribution
 - Dirichlet distribution is the conjugate prior of the multinomial distribution
 - Enables sum to one technique!

Gibbs Sampling Formula (1)

- $$P(W, Z; \alpha, \beta) = \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v) \Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v) \Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)} \prod_{j=1}^M \frac{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i) \Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i) \Gamma(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i)}$$
- Here, W , α and β are assumed or data-points, and Z is the target of sampling.
 - Gibbs sampling iterates the element of Z , one by one.
 - Therefore, we need to derive a formula of a single element Z when all other element of Z , W , α and β are given.
- $$P(Z_{(m,l)} = k | Z_{-(m,l)}, W; \alpha, \beta) = \frac{P(Z_{(m,l)} = k, Z_{-(m,l)}, W; \alpha, \beta)}{P(Z_{-(m,l)}, W; \alpha, \beta)} \\ \propto P(Z_{(m,l)} = k, Z_{-(m,l)}, W; \alpha, \beta)$$
 - $Z_{(m,l)}$ is the topic assignment on the l -th word of m -th document



Gibbs Sampling Formula (2)

- $$P(W, Z; \alpha, \beta) = \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v) \Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v) \Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)} \prod_{j=1}^M \frac{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i) \Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i) \Gamma(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i)}$$
 - $$= \left(\frac{\Gamma(\sum_{v=1}^V \beta_v)}{\prod_{v=1}^V \Gamma(\beta_v)} \right)^K \left(\frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \right)^M \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v)}{\Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)} \prod_{j=1}^M \frac{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i)}{\Gamma(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i)}$$
 - $$\propto \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v)}{\Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)} \prod_{j=1}^M \frac{\prod_{i=1}^K \Gamma(n_{j,(\cdot)}^i + \alpha_i)}{\Gamma(\sum_{i=1}^K n_{j,(\cdot)}^i + \alpha_i)}$$
- Now, apply that $P(Z_{(m,l)} = k, Z_{-(m,l)}, W; \alpha, \beta)$
 - $$\propto \prod_{i=1}^K \frac{\prod_{v=1}^V \Gamma(n_{(\cdot),v}^i + \beta_v)}{\Gamma(\sum_{v=1}^V n_{(\cdot),v}^i + \beta_v)} \times \frac{\prod_{i=1}^K \Gamma(n_{m,(\cdot)}^i + \alpha_i)}{\Gamma(\sum_{i=1}^K n_{m,(\cdot)}^i + \alpha_i)}: \text{Fixing document by } \mathbf{m}$$
 - $$\propto \prod_{i=1}^K \frac{\Gamma(n_{(\cdot),v}^i + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^i + \beta_r)} \times \frac{\prod_{i=1}^K \Gamma(n_{m,(\cdot)}^i + \alpha_i)}{\Gamma(\sum_{i=1}^K n_{m,(\cdot)}^i + \alpha_i)}: \text{Fixing word by } \mathbf{l}$$
 - $$\propto \prod_{i=1}^K \frac{\Gamma(n_{(\cdot),v}^i + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^i + \beta_r)} \times \prod_{i=1}^K \Gamma(n_{m,(\cdot)}^i + \alpha_i): \text{Remove a constant } \Gamma(\sum_{i=1}^K n_{m,(\cdot)}^i + \alpha_i)$$

Gibbs Sampling Formula (3)

- $P(Z_{(m,l)} = k | Z_{-(m,l)}, W; \alpha, \beta) \propto P(Z_{(m,l)} = k, Z_{-(m,l)}, W; \alpha, \beta)$
 - $\propto \prod_{i=1}^K \frac{\Gamma(n_{(\cdot),v}^i + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^i + \beta_r)} \times \prod_{i=1}^K \Gamma(n_{m,(\cdot)}^i + \alpha_k)$
- Now, we set $n_{j,r}^{i,-(m,l)}$ as $n_{j,r}^i$ excluding the count from the topic assignment of $Z_{(m,l)}$
 - $\propto \prod_{i=1, i \neq k}^K \frac{\Gamma(n_{(\cdot),v}^{i,-(m,n)} + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^{i,-(m,n)} + \beta_r)} \times \prod_{i=1, i \neq k}^K \Gamma(n_{m,(\cdot)}^{i,-(m,n)} + \alpha_k)$

$$\times \frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)} + \beta_v + 1)}{\Gamma((\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r) + 1)} \times \Gamma(n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k + 1)$$
 - Take out the k-th topic assignment because, the count of the k-th topic assignment count will be increased by 1 compared to $n_{(\cdot),(\cdot)}^{k,-(m,l)}$
 - Notice the increment of 1 in the separated multiplication
 - $\propto \prod_{i=1, i \neq k}^K \frac{\Gamma(n_{(\cdot),v}^{i,-(m,n)} + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^{i,-(m,n)} + \beta_r)} \times \prod_{i=1, i \neq k}^K \Gamma(n_{m,(\cdot)}^{i,-(m,n)} + \alpha_k)$

$$\times \frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)} + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r)} \times \Gamma(n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k) \times \frac{n_{(\cdot),v}^{k,-(m,n)} + \beta_v}{(\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r)} \times (n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k)$$
 - Definition of $\Gamma(x) = (x-1)!$
 - Therefore, $\Gamma(x+1) = (x)! \times x$

Gibbs Sampling Formula (4)

- $$P(Z_{(m,l)} = k | Z_{-(m,l)}, W; \alpha, \beta) \propto P(Z_{(m,l)} = k, Z_{-(m,l)}, W; \alpha, \beta)$$
 - $$\propto \prod_{i=1, i \neq k}^K \frac{\Gamma(n_{(\cdot),v}^{i,-(m,n)} + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^{i,-(m,n)} + \beta_r)} \times \prod_{i=1, i \neq k}^K \Gamma(n_{m,(\cdot)}^{i,-(m,n)} + \alpha_k)$$

$$\times \frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)} + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r)} \times \Gamma(n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k) \times \frac{n_{(\cdot),v}^{k,-(m,n)} + \beta_v}{(\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r)} \times (n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k)$$
 - $$\propto \prod_{i=1}^K \frac{\Gamma(n_{(\cdot),v}^{i,-(m,n)} + \beta_v)}{\Gamma(\sum_{r=1}^V n_{(\cdot),r}^{i,-(m,n)} + \beta_r)} \times \prod_{i=1}^K \Gamma(n_{m,(\cdot)}^{i,-(m,n)} + \alpha_k)$$

$$\times \frac{n_{(\cdot),v}^{k,-(m,n)} + \beta_v}{(\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r)} \times (n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k)$$
 - Absolved the $i=k$ case in to the large operator of multiplications
 - The topic assignment count used for the large multiplication is same to the assignment of any k on the word assignment of $Z_{(m,l)}$
 - Therefore, only the meaningful proportion is the separated single multiplications
- $$P(Z_{(m,l)} = k | Z_{-(m,l)}, W; \alpha, \beta) \propto \frac{n_{(\cdot),v}^{k,-(m,n)} + \beta_v}{(\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r)} \times (n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k)$$
 - Finally, this formula is simplified enough to calculate the likelihood of assigning k to $Z_{(m,l)}$
 - To become a probability, we need to normalize the above formula.

LDA Model Parameter Inference Routine with Gibbs

- LDA(TextCorpus T , α , β)
 - Randomly, initialize Z assignment on T
 - Count $n_{j,r}^i$ with the initial Z assignment
 - While performance measure (i.e. perplexity) converges
 - For $m = 1$ to T 's document number
 - For $l = 1$ to T_m 's document word length
 - Sampling k from $P(Z_{(m,l)} = k | Z_{-(m,l)}, W; \alpha, \beta) \propto \frac{n_{(\cdot),v}^{k,-(m,n)} + \beta_v}{(\sum_{r=1}^V n_{(\cdot),r}^{k,-(m,n)} + \beta_r)} \times (n_{m,(\cdot)}^{k,-(m,n)} + \alpha_k)$
 - Adjust $n_{j,r}^i$ by assigning $Z_{(m,l)} = k$
 - Calculate the most likely estimation on θ and φ
 - Return θ and φ
- Note
 - θ represents the document-topic probability
 - φ represents the topic-word probability
 - Perplexity is the measurement on the quality of the soft-clustering, and calculating it may take some time. Hence, many cases, we just set the iteration number.