Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:

Receptor A  $x_1$
Receptor B  $x_2$

Kinase C  $x_3$
Kinase D  $x_4$
Kinase E  $x_5$

TF F  $x_6$

Gene G  $x_7$
Gene H  $x_8$
What is a Graphical Model?
--- example from a signal transduction pathway

- A possible world for cellular signal transduction:

  - Receptor A $X_1$
  - Kinase C $X_3$
  - Kinase D $X_4$
  - Kinase E $X_5$
  - TF F $X_6$
  - Gene G $X_7$
  - Gene H $X_8$
GM: Structure Simplifies Representation

- Dependencies among variables

Receptor A \( x_1 \) → Kinase C \( x_3 \) → TF F \( x_6 \) → Gene G \( x_7 \)

Receptor B \( x_2 \) → Kinase D \( x_4 \) → Kinase E \( x_5 \) → Gene H \( x_8 \)

Membrane → Cytosol → Nucleus

© Eric Xing @ CMU, 2014
If $X_i$'s are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_2|X_1) P(X_4|X_2) P(X_5|X_2)$$

$$P(X_6|X_3, X_4) P(X_7|X_6) P(X_8|X_5, X_6)$$

Why we may favor a PGM?

- **Representation cost**: how many probability statements are needed?
  
  $2+2+4+4+8+4+8=36$, an 8-fold reduction from $2^8$!

- **Algorithms for systematic and efficient inference/learning computation**
  
  - Exploring the graph structure and probabilistic (e.g., Bayesian, Markovian) semantics

- **Incorporation of domain knowledge and causal (logical) structures**
There are two components to any GM:

- the \textit{qualitative} specification
- the \textit{quantitative} specification
Qualitative Specification

Where does the qualitative specification come from?

- Prior knowledge of causal relationships
- Prior knowledge of modular relationships
- Assessment from experts
- Learning from data
- We simply link a certain architecture (e.g. a layered graph)
- ...
Two types of GMs

- **Directed edges** give causality relationships (Bayesian Network or Directed Graphical Model):

\[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)
= P(X_1) P(X_2) P(X_3|X_1) P(X_4|X_2) P(X_5|X_2)
P(X_6|X_3, X_4) P(X_7|X_6) P(X_8|X_5, X_6)
\]

- **Undirected edges** simply give correlations between variables (Markov Random Field or Undirected Graphical model):

\[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)
= \frac{1}{Z} \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2)
+ E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}
Bayesian Network:

- A BN is a directed graph whose nodes represent the random variables and whose edges represent direct influence of one variable on another.

- It is a data structure that provides the skeleton for representing a joint distribution compactly in a factorized way;

- It offers a compact representation for a set of conditional independence assumptions about a distribution;

- We can view the graph as encoding a generative sampling process executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents. In other words, each variable is a stochastic function of its parents.
Bayesian Network: Factorization Theorem

Theorem:
Given a DAG, the most general form of the probability distribution that is consistent with the graph factors according to “node given its parents”:

\[ P(X) = \prod_{i=1}^{d} P(X_i | X_{\pi_i}) \]

where \( X_{\pi_i} \) is the set of parents of \( X_i \), \( d \) is the number of nodes (variables) in the graph.

\[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6) \]
Bayesian Network: Conditional Independence Semantics

Structure: DAG

- Meaning: a node is conditionally independent of every other node in the network outside its Markov blanket.

- Local conditional distributions (CPD) and the DAG completely determine the joint dist.

- Give causality relationships, and facilitate a generative process.
Graph separation criterion

- D-separation criterion for Bayesian networks (D for Directed edges):

  **Definition**: variables $x$ and $y$ are *D-separated* (conditionally independent) given $z$ if they are separated in the *moralized* ancestral graph

- Example:

  
  ![Original Graph](image1)
  ![Ancestral Graph](image2)
  ![Moral Ancestral Graph](image3)

© Eric Xing @ CMU, 2014
Global Markov properties of DAGs

• $X$ is d-separated (directed-separated) from $Z$ given $Y$ if we can't send a ball from any node in $X$ to any node in $Z$ using the "Bayes-ball" algorithm illustrated below (and plus some boundary conditions):

  • Defn: $I(G)$ = all independence properties that correspond to d-separation:

$$I(G) = \left\{ X \perp Z \mid Y : \text{dsep}_G(X; Z \mid Y) \right\}$$

• D-separation is sound and complete
Towards quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

- **The Equivalence Theorem**
  
  For a graph G,
  
  Let $\mathcal{D}_1$ denote the family of all distributions that satisfy $I(G)$,
  
  Let $\mathcal{D}_2$ denote the family of all distributions that factor according to $G$,
  
  Then $\mathcal{D}_1 \equiv \mathcal{D}_2$. 
Example

- Speech recognition

Hidden Markov Model
Knowledge Engineering

- Picking variables
  - Observed
  - Hidden

- Picking structure
  - CAUSAL
  - Generative

- Picking Probabilities
  - Zero probabilities
  - Orders of magnitudes
  - Relative values
Example, con'd

- Evolution

Tree Model
Conditional probability tables (CPTs)

\[
P(a, b, c, d) = P(a)P(b)P(c | a, b)P(d | c)
\]

<table>
<thead>
<tr>
<th>a⁰</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>a¹</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b⁰</th>
<th>0.33</th>
</tr>
</thead>
<tbody>
<tr>
<td>b¹</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c⁰</th>
<th>0.45</th>
<th>1</th>
<th>0.9</th>
<th>0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>c¹</td>
<td>0.55</td>
<td>0</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d⁰</th>
<th>0.3</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>d¹</td>
<td>0.7</td>
<td>0.5</td>
</tr>
</tbody>
</table>

© Eric Xing @ CMU, 2014
Conditional probability density func. (CPDs)

\[ P(a,b,c,d) = P(a)P(b)P(c|a,b)P(d|c) \]

\[ A \sim N(\mu_a, \Sigma_a) \quad B \sim N(\mu_b, \Sigma_b) \]

\[ C \sim N(A+B, \Sigma_c) \]

\[ D \sim N(\mu_a+C, \Sigma_a) \]
Conditionally Independent Observations

Data

Model parameters

$\theta$

$y_1$, $y_2$, $\ldots$, $y_{n-1}$, $y_n$
“Plate” Notation

Plate = rectangle in graphical model

variables within a plate are replicated in a conditionally independent manner

Model parameters

Data = \{y_1, \ldots, y_n\}
Example: Gaussian Model

Generative model:

$$p(y_1, \ldots, y_n \mid \mu, \sigma) = P \ p(y_i \mid \mu, \sigma)$$

$$= p(\text{data} \mid \text{parameters})$$

$$= p(D \mid \theta)$$

where $\theta = \{\mu, \sigma\}$

- Likelihood  $$= p(\text{data} \mid \text{parameters})$$
  $$= p(D \mid \theta)$$
  $$= L(\theta)$$

- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
  - Often easier to work with $\log L(\theta)$
Example: Bayesian Gaussian Model

\[ y_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \]

Note: priors and parameters are assumed independent here

© Eric Xing @ CMU, 2014
Markov Random Fields

Structure: an *undirected graph*

- Meaning: a node is *conditionally independent* of every other node in the network given its *Directed neighbors*

- Local contingency functions (*potentials*) and the *cliques* in the graph completely determine the *joint dist.*

- Give *correlations* between variables, but no explicit way to generate samples
Global Markov property

Let $H$ be an undirected graph:

- $B$ separates $A$ and $C$ if every path from a node in $A$ to a node in $C$ passes through a node in $B$: $\text{sep}_H(A; C | B)$

- A probability distribution satisfies the **global Markov property** if for any disjoint $A, B, C$, such that $B$ separates $A$ and $C$, $A$ is independent of $C$ given $B$: $I(H) = \left\{ A \perp C | B : \text{sep}_H(A; C | B) \right\}$
Representation

- Defn: an undirected graphical model represents a distribution \( P(X_1, \ldots, X_n) \) defined by an undirected graph \( H \), and a set of positive potential functions \( y_c \) associated with cliques of \( H \), s.t.

\[
P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)
\]

where \( Z \) is known as the partition function:

\[
Z = \sum_{x_1, \ldots, x_n} \prod_{c \in C} \psi_c(x_c)
\]

- Also known as Markov Random Fields, Markov networks ...

- The potential function can be understood as an contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.
Cliques

- For $G = \{V,E\}$, a complete subgraph (clique) is a subgraph $G' = \{V',E'\}$ such that nodes in $V'$ are fully interconnected.
- A (maximal) clique is a complete subgraph s.t. any superset $V'' \supseteq V'$ is not complete.
- A sub-clique is a not-necessarily-maximal clique.

**Example:**
- max-cliques = \{A,B,D\}, \{B,C,D\},
- sub-cliques = \{A,B\}, \{C,D\}, \rightarrow all edges and singletons
Example UGM – using max cliques

For discrete nodes, we can represent $P(x_1, x_2, x_3, x_4)$ as two 3D tables instead of one 4D table.

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi_c(x_{124}) \times \psi_c(x_{234})$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_c(x_{124}) \times \psi_c(x_{234})$$

- For discrete nodes, we can represent $P(X_{1:4})$ as two 3D tables instead of one 4D table.
Example UGM – using subcliques

For discrete nodes, we can represent $P(X_{1:4})$ as 5 2D tables instead of one 4D table.
Exponential Form

- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential \( \psi_c(x_c) \) in an unconstrained form using a real-value "energy" function \( \phi_c(x_c) \):

\[
\psi_c(x_c) = \exp\left\{-\phi_c(x_c)\right\}
\]

For convenience, we will call \( \phi_c(x_c) \) a potential when no confusion arises from the context.

- This gives the joint a nice additive structure

\[
p(x) = \frac{1}{Z} \exp\left\{-\sum_{c \in C} \phi_c(x_c)\right\} = \frac{1}{Z} \exp\left\{-H(x)\right\}
\]

where the sum in the exponent is called the "free energy":

\[
H(x) = \sum_{c \in C} \phi_c(x_c)
\]

- In physics, this is called the "Boltzmann distribution".
- In statistics, this is called a log-linear model.
A fully connected graph with pairwise (edge) potentials on binary-valued nodes (for $x_i \in \{-1,+1\}$ or $x_i \in \{0,1\}$) is called a Boltzmann machine.

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \exp \left\{ \sum_{ij} \phi_{ij}(x_i, x_j) \right\}$$

$$= \frac{1}{Z} \exp \left\{ \sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C \right\}$$

Hence the overall energy function has the form:

$$H(x) = \sum_{ij} (x_i - \mu) \Theta_{ij} (x_j - \mu) = (x - \mu)^T \Theta (x - \mu)$$
Example: Ising (spin-glass) models

- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors.

- Same as sparse Boltzmann machine, where \( \theta_{ij} \neq 0 \) iff \( i, j \) are neighbors.
  - e.g., nodes are pixels, potential function encourages nearby pixels to have similar intensities.

- Potts model: multi-state Ising model.
Example: Modeling Go

This is the middle position of a Go game. Overlaid is the estimate for the probability of becoming black or white for every intersection. Large squares mean the probability is higher.
An (incomplete) genealogy of graphical models

(Picture by Zoubin Ghahramani and Sam Roweis)
Advanced Introduction to Machine Learning

Markov Chain Monte Carlo

Eric Xing

Lecture 14, October 20, 2014

Reading:
© Eric Xing @ CMU, 2014
Approaches to inference

- **Exact inference algorithms**
  - The elimination algorithm
  - Belief propagation
  - The junction tree algorithms (but will not cover in detail here)

- **Approximate inference techniques**
  - Variational algorithms
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods
Monte Carlo methods

- Draw random samples from the desired distribution
- Yield a stochastic representation of a complex distribution
  - marginals and other expectations can be approximated using sample-based averages
    \[
    E[f(x)] = \frac{1}{N} \sum_{t=1}^{N} f(x^{(t)})
    \]
- **Asymptotically** exact and easy to apply to arbitrary models
- Challenges:
  - how to draw samples from a given dist. (not all distributions can be trivially sampled)?
  - how to make better use of the samples (not all samples are useful, or equally useful, see an example later)?
  - how to know we've sampled enough?
Example: naive sampling

- Construct samples according to probabilities given in a BN.

**Alarm example: (Choose the right sampling sequence)**

1) **Sampling:** \( P(B) = \langle 0.001, 0.999 \rangle \) suppose it is false, \( B_0 \). Same for \( E_0 \). \( P(A|B_0, E_0) = \langle 0.001, 0.999 \rangle \) suppose it is false...

2) **Frequency counting:** In the samples right, \( P(J|A_0) = P(J,A_0)/P(A_0) = \langle 1/9, 8/9 \rangle \).

<table>
<thead>
<tr>
<th>( E_0 )</th>
<th>( B_0 )</th>
<th>( A_0 )</th>
<th>( M_0 )</th>
<th>( J_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_1 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
<tr>
<td>( E_0 )</td>
<td>( B_0 )</td>
<td>( A_0 )</td>
<td>( M_0 )</td>
<td>( J_0 )</td>
</tr>
</tbody>
</table>
Example: naive sampling

- Construct samples according to probabilities given in a BN.

Alarm example: (Choose the right sampling sequence)

3) what if we want to compute $P(J|A1)$? we have only one sample ...

4) what if we want to compute $P(J|B1)$? No such sample available!
$P(J|A1)=P(J,B1)/P(B1)$ can not be defined.

For a model with hundreds or more variables, rare events will be very hard to garner enough samples even after a long time or sampling ...
Monte Carlo methods (cond.)

- **Direct Sampling**
  - We have seen it.
  - Very difficult to populate a high-dimensional state space

- **Rejection Sampling**
  - Create samples like direct sampling, only count samples which is consistent with given evidences.

- **Likelihood weighting, ...**
  - Sample variables and calculate evidence weight. Only create the samples which support the evidences.

- **Markov chain Monte Carlo (MCMC)**
  - Metropolis-Hasting
  - Gibbs
Construct a Markov chain whose stationary distribution is the target density \( P(X|e) \).

Run for \( T \) samples (burn-in time) until the chain converges/mixes/reaches stationary distribution.

Then collect \( M \) (correlated) samples \( x_m \).

Key issues:
- Designing proposals so that the chain mixes rapidly.
- Diagnosing convergence.
Markov Chains

- **Definition:**
  - Given an n-dimensional state space
  - Random vector $X = (x_1, \ldots, x_n)$
  - $x^{(t)} = x$ at time-step $t$
  - $x^{(t)}$ transitions to $x^{(t+1)}$ with prob
    $P(x^{(t+1)} | x^{(t)}, \ldots, x^{(1)}) = T(x^{(t+1)} | x^{(t)}) = T(x^{(t)} \rightarrow x^{(t+1)})$

- **Homogenous:** chain determined by state $x^{(0)}$, fixed *transition kernel* $T$ (rows sum to 1)

- **Equilibrium:** $\pi(x)$ is a *stationary (equilibrium) distribution* if
  $\pi(x') = \sum_x \pi(x) T(x \Rightarrow x')$.
  i.e., is a left eigenvector of the transition matrix $\pi^T T = \pi^T$.

![Diagram showing transitions between states X1, X2, X3 with probabilities]
Markov Chains

- An MC is **irreducible** if the transition graph is connected.
- An MC is **aperiodic** if it is not trapped in cycles.
- An MC is **ergodic** (regular) if you can get from state $x$ to $x'$ in a finite number of steps.

**Detailed balance:**

\[
\text{prob}(x^{(t)} \rightarrow x^{(i-1)}) = \text{prob}(x^{(t-1)} \rightarrow x^{(t)})
\]

\[
p(x^{(t)}) \mathbb{T}(x^{(t-1)} | x^{(t)}) = p(x^{(t-1)}) \mathbb{T}(x^{(t-1)} | x^{(t-1)})
\]

summing over $x^{(t-1)}$

\[
p(x^{(t)}) = \sum_{x^{(t-1)}} p(x^{(t-1)}) \mathbb{T}(x^{(t)} | x^{(t-1)})
\]

- Detailed bal $\rightarrow$ stationary dist exists.
Metropolis-Hastings

- Treat the target distribution as stationary distribution
- Sample from an easier proposal distribution, followed by an acceptance test
- This induces a transition matrix that satisfies detailed balance

  - MH proposes moves according to $Q(x' | x)$ and accepts samples with probability $A(x' | x)$.
  - The induced transition matrix is $T(x \rightarrow x') = Q(x' | x)A(x' | x)$
  - Detailed balance means
    \[
    \pi(x)Q(x' | x)A(x' | x) = \pi(x')Q(x | x')A(x | x')
    \]
  - Hence the acceptance ratio is
    \[
    A(x' | x) = \min \left(1, \frac{\pi(x')Q(x | x')}{\pi(x)Q(x' | x)} \right)
    \]
Metropolis-Hastings

1. Initialize $x^{(0)}$

2. While not mixing  // burn-in
   - $x = x^{(t)}$
   - $t += 1$
   - sample $u \sim \text{Unif}(0,1)$
   - sample $x^* \sim Q(x^* | x)$
     - if $u < A(x^* | x) = \min \left( 1, \frac{\pi(x^*) Q(x | x^*)}{\pi(x) Q(x^* | x)} \right)$
       - $x^{(t)} = x^*$  // transition
     - else
       - $x^{(t)} = x$  // stay in current state

- Reset $t=0$, for $t = 1:N$
  - $x(t+1) \leftarrow \text{Draw sample } (x(t))$

Function
Draw sample $(x(t))$
MCMC example

\[ q(x^*|x) \sim N(x^{(i)}, 100) \]

\[ p(x) \sim 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2) \]
Gibbs sampling

- Gibbs sampling is an MCMC algorithm that is especially appropriate for inference in graphical models.

- The procedure
  - we have variable set $X = \{x_1, x_2, x_3, \ldots, x_N\}$ for a GM
  - at each step one of the variables $X_i$ is selected (at random or according to some fixed sequences), denote the remaining variables as $X_{-i}$, and its current value as $x_i^{(t-1)}$
    - Using the "alarm network" as an example, say at time $t$ we choose $X_E$, and we denote the current value assignments of the remaining variables, $X_{-E}$, obtained from previous samples, as $x_{-E}^{(t-1)} = \{x_B^{(t-1)}, x_A^{(t-1)}, x_J^{(t-1)}, x_M^{(t-1)}\}$
  - the conditional distribution $p(X_i | x_{-i}^{(t-1)})$ is computed
  - a value $x_i^{(t)}$ is sampled from this distribution
  - the sample $x_i^{(t)}$ replaces the previous sampled value of $X_i$ in $X$.
    - i.e., $X^{(t)} = X_{-E}^{(t-1)} \cup x_i^{(t)}$
Markov Blanket

- **Markov Blanket in BN**
  - A variable is independent from others, given its parents, children and children’s parents (d-separation).

- **MB in MRF**
  - A variable is independent all its non-neighbors, given all its direct neighbors.

\[ p(X_i | X_{\neg i}) = p(X_i | MB(X_i)) \]

- **Gibbs sampling**
  - Every step, choose one variable and sample it by \( P(X|MB(X)) \) based on previous sample.
Gibbs sampling of the alarm network

To calculate $P(J|B_1,M_1)$

Choose $(B_1,E_0,A_1,M_1,J_1)$ as a start

Evidences are $B_1$, $M_1$, variables are $A$, $E$, $J$.

Choose next variable as $A$

Sample $A$ by $P(A|MB(A)) = P(A|B_1,E_0,M_1,J_1)$ suppose to be false.

$(B_1, E_0, A_0, M_1, J_1)$

Choose next random variable as $E$, sample $E \sim P(E|B_1,A_0)$

- $MB(A) = \{B, E, J, M\}$
- $MB(E) = \{A, B\}$
Example

First 100 iterations of sample3

© Eric Xing @ CMU, 2014
Example:
Example
Example

\[ P(J_1 \mid B_1, M_1) = 0.90 \]
\[ P(J_1 \mid E_1, M_0) = 0.14 \]
\[ P(E_1 \mid J_1) = 0.01 \]
\[ P(E_1 \mid M_1) = 0.04 \]
\[ P(E_1 \mid M_1, J_1) = 0.17 \]

Gibbs sampling of alarm network
Gibbs sampling

- Gibbs sampling is a special case of MH
- The transition matrix updates each node one at a time using the following proposal:
  \[ Q((x_i, x_{-i}) \rightarrow (x_i', x_{-i})) = p(x_i'| x_{-i}) \]
- This is efficient since for two reasons
  - It leads to samples that is always accepted
    \[
    A((x_i, x_{-i}) \rightarrow (x_i', x_{-i})) = \min \left( 1, \frac{p(x_i', x_{-i})Q((x_i', x_{-i}) \rightarrow (x_i, x_{-i}))}{p(x_i, x_{-i})Q((x_i, x_{-i}) \rightarrow (x_i', x_{-i}))} \right) 
    = \min \left( 1, \frac{p(x_i'| x_{-i})p(x_{-i})p(x_i | x_{-i})}{p(x_i | x_{-i})p(x_{-i})p(x_i'| x_{-i})} \right) = \min(1,1)
    \]
    Thus
    \[ T((x_i, x_{-i}) \rightarrow (x_i', x_{-i})) = p(x_i'| x_{-i}) \]
  - It is efficient since \( p(x_i'| x_{-i}) \) only depends on the values in \( X_i \)'s Markov blanket
Gibbs sampling

- Scheduling and ordering:
  - Sequential sweeping: in each "epoch" $t$, touch every r.v. in some order and yield an new sample, $x^{(t)}$, after every r.v. is resampled
  - Randomly pick an r.v. at each time step

- Blocking:
  - Large state space: state vector $X$ comprised of many components (high dimension)
  - Some components can be correlated and we can sample components (i.e., subsets of r.v.,) one at a time

- Gibbs sampling can fail if there are deterministic constraint
  - Suppose we observe $Z = 1$. The posterior has 2 modes: $P(X = 1, Y = 0|Z = 1)$ and $P(X = 0, Y = 1|Z = 1)$. if we start in mode 1, $P(X|Y = 0, Z = 1)$ leaves $X = 1$, so we can’t move to mode 2 (Reducible Markov chain).
  - If all states have non-zero probability, the MC is guaranteed to be regular.
  - Sampling blocks of variables at a time can help improve mixing.
GOOD! Chains

© Eric Xing @ CMU, 2014
Chains

Met. with Proposal Unif(x-0.1,x+0.1)

The wall

No mixing yet!!

n=3, alpha=1, m=0.92, s=1.55, Nmet=5.
The Art of simulation

- Run several chains
- Start at over-dispersed points
- Monitor the log lik.
- Monitor the serial correlations
- Monitor acceptance ratios
- Re-parameterize (to get approx. indep.)
- Re-block (Gibbs)
- Collapse (int. over other pars.)
- Run with troubled pars. fixed at reasonable vals.
Summary

- Random walk through state space
- Can simulate multiple chains in parallel
- Much hinges on proposal distribution \( Q \)
  - Want to visit state space where \( p(X) \) puts mass
  - Want \( A(x^*|x) \) high in modes of \( p(X) \)
  - Chain mixes well
- Convergence diagnosis
  - How can we tell when burn-in is over?
  - Run multiple chains from different starting conditions, wait until they start “behaving similarly”.
  - Various heuristics have been proposed.
Intro to Topic Models

Eric Xing

Lecture 15, October 20, 2014

Reading: Tutorial on Topic Model @ ACL12
We are inundated with data …

- Humans cannot afford to deal with (e.g., search, browse, or measure similarity) a huge number of text and media documents.
- We need computers to help out …
A task:

- Say, we want to have a mapping ..., so that

- Compare similarity
- Classify contents
- Cluster/group/categorize docs
- Distill semantics and perspectives
- ..
Representation:

- **Data:** Bag of Words Representation

As for the Arabian and Palestinian voices that are against the current negotiations and the so-called peace process, they are not against peace per se, but rather for their well-founded predictions that Israel would NOT give an inch of the West Bank (and most probably the same for Golan Heights) back to the Arabs. An 18 months of "negotiations" in Madrid, and Washington proved these predictions. Now many will jump on me saying why are you blaming Israelis for no-result negotiations. I would say why would the Arabs stall the negotiations, what do they have to loose?

- Each document is a vector in the word space
- Ignore the order of words in a document. Only count matters!
- A high-dimensional and sparse representation
  - Not efficient text processing tasks, e.g., search, document classification, or similarity measure
  - Not effective for browsing
Latent Semantic Structure in GM

Distribution over words

\[ P(w) = \sum_{\ell} P(w, \ell) \]

Inferring latent structure

\[ P(\ell \mid w) = \frac{P(w \mid \ell)P(\ell)}{P(w)} \]
How to Model Semantics?

Q: What is it about?
A: Mainly MT, with syntax, some learning

A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

Source Target SMT Alignment Score BLEU

Parse Tree Noun Phrase Grammar CFG

likelihood EM Hidden Parameters Estimation argMax

Unigram over vocabulary

AdMixing Proportion

© Eric Xing @ CMU, 2014
Why this is Useful?

- Q: What is it about?
- A: Mainly MT, with syntax, some learning

<table>
<thead>
<tr>
<th></th>
<th>MT</th>
<th>Syntax</th>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- Q: give me similar document?
  - Structured way of browsing the collection

- Other tasks
  - Dimensionality reduction
    - TF-IDF vs. topic mixing proportion
    - Classification, clustering, and more …

**A Hierarchical Phrase-Based Model for Statistical Machine Translation**

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.
Words in Contexts

- “It was a nice **shot**.”
Words in Contexts (con'd)

- the opposition Labor Party fared even worse, with a predicted 35 seats, seven less than last election.
A possible generative process of a document

**Mixture Components (distributions over elements)**

**admixing weight vector \( \theta \)** (represents all components’ contributions)

**Bayesian approach:** use priors

Admixture weights \( \sim \text{Dirichlet}(\alpha) \)

Mixture components \( \sim \text{Dirichlet}(\Gamma) \)

**DOCUMENT 1:** 
- money\(^1\) bank\(^1\) bank\(^1\) loan\(^1\) river\(^2\) stream\(^2\)
- bank\(^1\) money\(^1\) river\(^2\) bank\(^1\) money\(^1\) bank\(^1\) loan\(^1\) money\(^1\) stream\(^2\) bank\(^1\) money\(^1\) bank\(^1\) bank\(^1\) loan\(^1\) river\(^2\) stream\(^2\)
- bank\(^1\) money\(^1\) river\(^2\) bank\(^1\) money\(^1\) bank\(^1\) bank\(^1\) loan\(^1\) bank\(^1\) money\(^1\) stream\(^2\)

**DOCUMENT 2:** 
- river\(^2\) stream\(^2\) bank\(^2\) stream\(^2\) bank\(^2\) money\(^1\) loan\(^1\) river\(^2\) stream\(^2\) loan\(^1\) bank\(^2\) river\(^2\) bank\(^2\)
- bank\(^1\) stream\(^2\) river\(^2\) loan\(^1\) bank\(^2\) stream\(^2\) bank\(^2\) money\(^1\) loan\(^1\) river\(^2\) stream\(^2\) bank\(^2\) stream\(^2\) bank\(^2\) money\(^1\) river\(^2\) stream\(^2\) loan\(^1\) bank\(^2\) river\(^2\) bank\(^2\) money\(^1\) bank\(^1\) stream\(^2\) bank\(^2\) stream\(^2\) bank\(^2\) money\(^1\)
Generating a document

- Draw $\theta$ from the prior
  For each word $n$
  - Draw $z_n$ from $\text{multinomial}(\theta)$
  - Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $\text{multinomial}(\beta_{z_n})$

Which prior to use?
Essentially a Bayesian pLSI:

\[ p(w) = \sum_z \int p(\theta) p(\beta) \left( \prod_{n=1}^{N} p(z_n|\theta) p(w_n|\beta_{z_n}) \right) d\theta d\beta \]

\[ \theta \sim \text{Dir}(\alpha) \]

\[ z_n \sim \text{Mult}(\theta) \]

\[ w_n \sim p(w_n|z_n, \beta) \]
Outcomes from a topic model

- The “topics” \( \beta \) in a corpus:

  ![Table of topics]

  - There is no name for each “topic”, you need to name it!
  - There is no objective measure of good/bad
  - The shown topics are the “good” ones, there are many many trivial ones, meaningless ones, redundant ones, … you need to manually prune the results
  - How many topics? …
Outcomes from a topic model

- The “topic vector” $\theta$ of each doc

- Create an embedding of docs in a “topic space”
- Their no ground truth of $\theta$ to measure quality of inference
- But on $\theta$ it is possible to define an “objective” measure of goodness, such as classification error, retrieval of similar docs, clustering, etc., of documents
- But there is no consensus on whether these tasks bear the true value of topic models …
Outcomes from a topic model

- The per-word topic indicator $z$:

  The William Randolph Hearst Foundation will give $1.25$ million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Not very useful under the bag of word representation, because of loss of ordering
- But it is possible to define simple probabilistic linguistic constraints (e.g, bi-grams) over $z$ and get potentially interesting results [Griffiths, Steyvers, Blei, & Tenenbaum, 2004]
Outcomes from a topic model

- Topic change trends

![Graphs showing topic change trends in "Theoretical Physics" and "Neuroscience" over time.](image)

[David Blei, MLSS09]

© Eric Xing @ CMU, 2014
The Big Picture

Unstructured Collection

Structured Topic Network

Word Simplex

Topic Discovery

Dimensionality Reduction

Topic Space
(e.g., a Simplex)

© Eric Xing @ CMU, 2014
Computation on LDA

- **Inference**
  - Given a Document $D$
  - Posterior: $P(\Theta | \mu, \Sigma, \beta, D)$
  - Evaluation: $P(D | \mu, \Sigma, \beta)$

- **Learning**
  - Given a collection of documents $\{D_i\}$
  - Parameter estimation

$$ \arg \max_{\mu, \Sigma, \beta} \sum \log(P(D_i | \mu, \Sigma, \beta)) $$
Exact Bayesian inference on LDA is intractable

- A possible query:

\[ p(\theta_n \mid D) = ? \]
\[ p(z_{n,m} \mid D) = ? \]

- Close form solution?

\[
p(\theta_n \mid D) = \frac{p(\theta_n, D)}{p(D)} \\
= \sum {\prod_n} \left( \prod_m p(x_{n,m} \mid \theta_{z_n}) p(z_{n,m} \mid \theta_n) \right) p(\theta_n \mid \alpha) p(\phi \mid G) d\theta_n \ d\beta \\
= \frac{p(D)}{p(D)} \\
p(D) = \sum {\prod_n} \left( \prod_m p(x_{n,m} \mid \theta_{z_n}) p(z_{n,m} \mid \theta_n) \right) p(\theta_n \mid \alpha) p(\beta \mid G) d\theta_1 \ldots d\theta_N d\beta
\]

- Sum in the denominator over \( T^m \) terms, and integrate over \( n \) \( k \)-dimensional topic vectors
Approximate Inference

● Variational Inference
  ● Mean field approximation (Blei et al)
  ● Expectation propagation (Minka et al)
  ● Variational 2\textsuperscript{nd}-order Taylor approximation (Ahmed and Xing)

● Markov Chain Monte Carlo
  ● Gibbs sampling (Griffiths et al)
Collapsed Gibbs sampling
(Tom Griffiths & Mark Steyvers)

- Collapsed Gibbs sampling
  - Integrate out $\theta$

For variables $z = z_1, z_2, \ldots, z_n$

Draw $z_i^{(t+1)}$ from $P(z_i|z_{-i}, w)$

$z_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_n^{(t)}$

\[
\{z^{(1)}, z^{(2)}, \ldots, z^{(T)}\}
\]

\[
\theta = \frac{1}{T} \sum_t z^{(t)}
\]
Gibbs sampling

- Need full conditional distributions for variable
- Since we only sample \( z \) we need

\[
P(z_i = j | z_{-i}, w) \propto P(w_i | z_i = j, z_{-i}, w_{-i}) P(z_i = j | z_{-i})
\]

\[
= \frac{n_{-i,j}^{(w_i)} + G}{n_{-i,j}^{(w)} + W} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d)} + T\alpha}
\]

- \( n_j^{(w)} \) number of times word \( w \) assigned to topic \( j \)
- \( n_j^{(d)} \) number of times topic \( j \) used in document \( d \)
## Gibbs sampling

### Iteration 1

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
**Gibbs sampling**

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

© Eric Xing @ CMU, 2014
### Gibbs sampling

**Iteration**

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)}}{n_{i,j}^{(-)}} + \frac{G}{W} \frac{n_{-i,j}^{(d_i)}}{n_{i,j}^{(-)}} + \alpha \]
### Gibbs sampling

**Table:**

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Equation:**

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{i,j}^{(w_i)} + G}{n_{-i,j}^{(w_i)}} + W \frac{n_{i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + T \alpha}$$

© Eric Xing @ CMU, 2014
Gibbs sampling

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + G}{n_{-i,j}^{(1)} + W G} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + T \alpha}$$

© Eric Xing @ CMU, 2014 86
Gibbs sampling

\[ P(z_i = j | z_{-i}, w) \propto \frac{n^{(w_i)}_{-i,j} + G}{n^{(i)}_{-i,j} + WG} \frac{n^{(d_i)}_{-i,j} + \alpha}{n_{-i,j} + T\alpha} \]

<table>
<thead>
<tr>
<th>(i)</th>
<th>(w_i)</th>
<th>(d_i)</th>
<th>(z_i)</th>
<th>(z_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Gibbs sampling

### Iteration 1 and 2

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + G}{n_{-i,j}^{(i)} + WG} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(\cdot)} + T\alpha}$$

© Eric Xing @ CMU, 2014
### Gibbs Sampling

**Iteration**

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>$w_i$</td>
<td>$d_i$</td>
<td>$z_i$</td>
<td>$z_i$</td>
</tr>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td>.</td>
</tr>
</tbody>
</table>

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + G}{n_{-i,j}^{(w_i)} + W G} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + T \alpha}$$
Gibbs sampling

<table>
<thead>
<tr>
<th>iteration</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>$w_i$</td>
<td>$d_i$</td>
<td>$z_i$</td>
<td>$z_i$</td>
</tr>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

$\theta = \frac{1}{T} \sum_t z^{(t)}$

$$P(z_i = j | z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + G}{n_{-i,j}^{(\cdot)} + WG \alpha} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(\cdot)} + WG \alpha}$$
Learning a TM

- Maximum likelihood estimation:

\[ \{\beta_1, \beta_2, \ldots, \beta_K\}, \alpha = \arg \max_{(\alpha, \beta)} \sum \log(P(D_i|\alpha, \beta)) \]

- Need statistics on topic-specific word assignment (due to \( z \)), topic vector distribution (due to \( \theta \)), etc.
  - E.g., this is the formula for topic \( k \):

\[
\beta_k = \frac{1}{\sum_d N_d} \sum_{d=1}^D \sum_{d_n=1}^{N_d} \delta(z_{d,n}, k) w_{d,d_n}
\]

- These are hidden variables, therefore need an EM algorithm (also known as data augmentation, or DA, in Monte Carlo paradigm)

- This is a “reduce” step in parallel implementation
Conclusion

- GM-based topic models are cool
  - Flexible
  - Modular
  - Interactive
- There are many ways of implementing topic models
  - unsupervised
  - supervised
- Efficient Inference/learning algorithms
  - GMF, with Laplace approx. for non-conjugate dist.
  - MCMC
- Many applications
  - ...
  - Word-sense disambiguation
  - Image understanding
  - Network inference