

TTIC 31210:  
Advanced Natural Language Processing

Kevin Gimpel  
Spring 2017

Lecture 13:  
Introduction to Bayesian  
Nonparametrics for NLP

# “Nonparametric”?

- nonparametric does not mean “no parameters”
- it means that “the number of parameters grows as the data grows”
- for our purposes, think of it as “some component of the model permits an unbounded set of something”

# Parametric or Nonparametric?

Model

Parametric or  
Nonparametric?

\*parametric if vocab fixed

# Parametric or Nonparametric?

Model	Parametric or Nonparametric?
Gaussian Mixture Model (GMM)	parametric
Hidden Markov Model (with GMM emissions)	parametric
Hidden Markov Model (for part-of-speech tagging, with multinomial emissions)	nonparametric*
n-gram language models	nonparametric
LDA	nonparametric*
LSTM language model	nonparametric*
character-level LSTM language model	parametric (assuming fixed set of characters)

\*parametric if vocab fixed

- “nonparametric modeling” in terms of vocab has a lot of engineering solutions:
  - use UNK for unknown words, do smoothing of high-order  $n$ -grams, etc.
- in this case, unbounded part of model is mostly determined by observed data, heuristics are useful
- modeling gets more interesting when unbounded part of model relates to **latent variables**

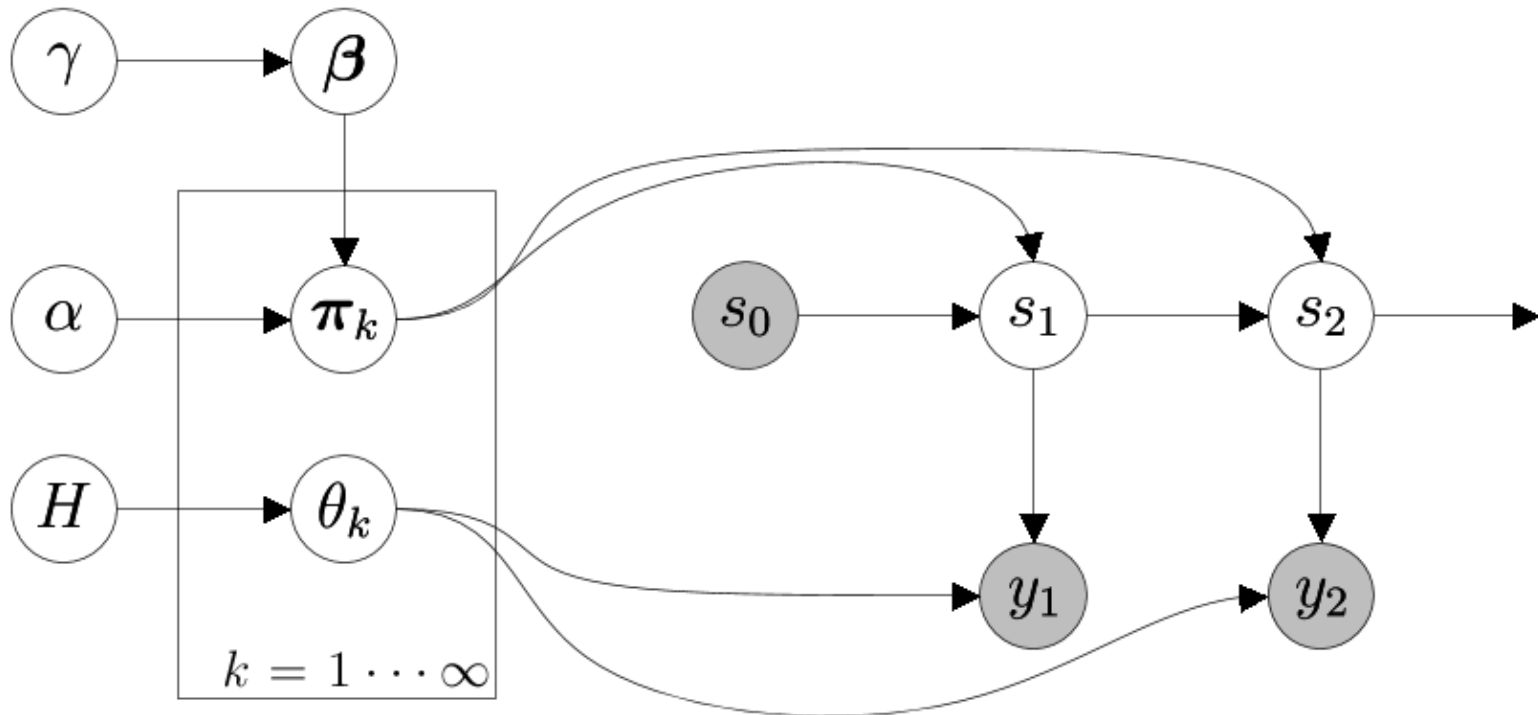
- when might you want to permit an unbounded set of latent items in a model?

# Infinite Mixture Model

- number of mixture components is unbounded (grows depending on the data)
- e.g., LDA with an unbounded set of topics

# “Infinite” HMM

- HMMs permit infinite sequences already
- what’s new here?
- infinite number of **hidden states**:





# “Infinite” PCFG

- PCFGs can already handle infinite-length derivations
- “infinite” here means an infinite number of **nonterminals**:

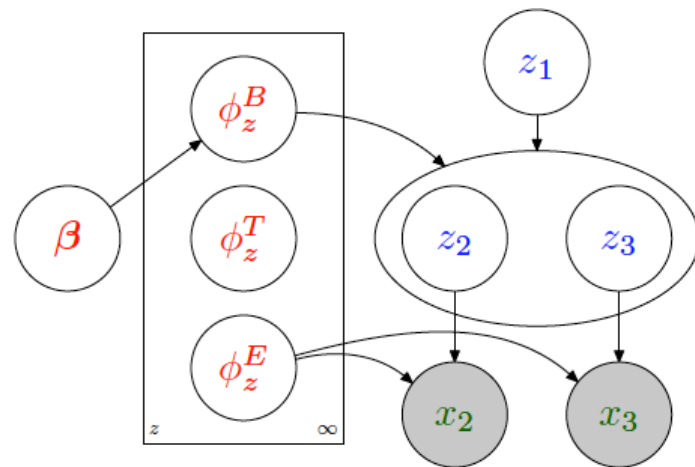
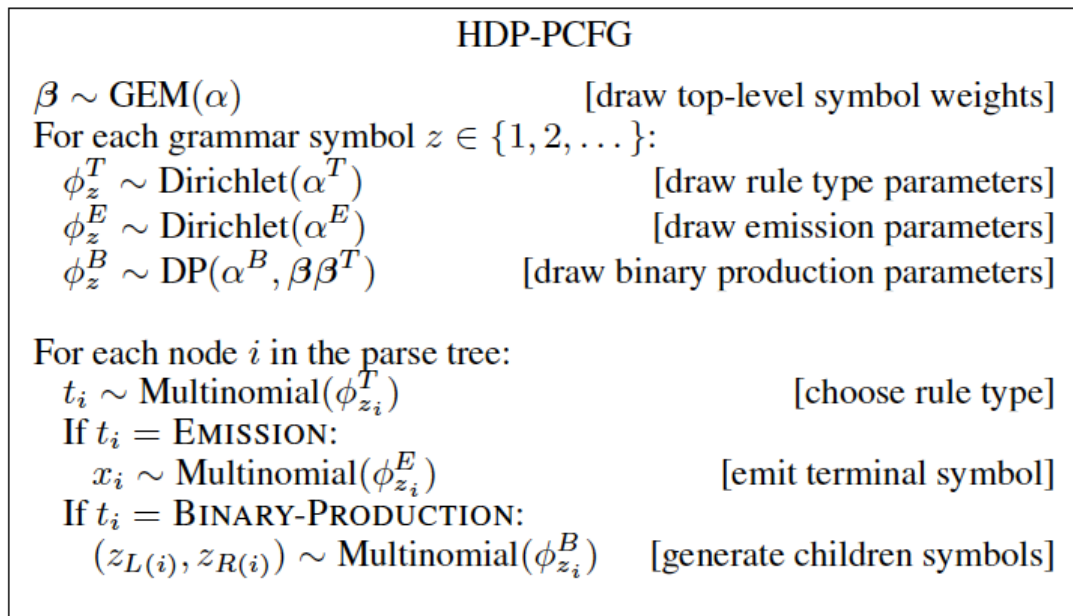


Figure 2: The definition and graphical model of the HDP-PCFG. Since parse trees have unknown structure, there is no convenient way of representing them in the visual language of traditional graphical models. Instead, we show a simple fixed example tree. Node 1 has two children, 2 and 3, each of which has one observed terminal child. We use  $L(i)$  and  $R(i)$  to denote the left and right children of node  $i$ .

- when might you want to permit an unbounded set of latent items in a model?
  - number of topics in LDA
  - number of Gaussians in a Gaussian Mixture Model
  - number of hidden states in an HMM
  - number of nonterminals in a PCFG
  - morph lexicon for morphological segmentation
  - lexicon for Chinese word segmentation
  - number of coreference chains in coreference resolution
  - number of dimensions in an embedding (?!)


- we need priors over distributions that permit an unbounded set of items

# LDA Generative Story

- 1: For each topic  $k = 1 \dots K$ , draw multinomial word distribution  $\beta_k \sim \text{Dirichlet}(\psi)$
- 2: For each document  $i$ :
  - a: Draw a multinomial topic distribution  $\theta^{(i)} \sim \text{Dirichlet}(\alpha)$
  - b: For each position  $j$  in document  $i$ :
    - i: Draw a topic  $z^{(i,j)} \sim \text{Multinomial}(\theta^{(i)})$
    - ii: Draw a word  $w^{(i,j)} \sim \text{Multinomial}(\beta_{z^{(i,j)}})$

$K = \# \text{ topics}$   
 $N = \# \text{ documents}$   
 $M = \# \text{ words in each document}$   
 $V = \# \text{ words in vocabulary}$

dimensionality of alpha  
must be  $K$  (the number of topics)



# Dirichlet Process (DP)

- “distribution over distributions”
- unlike Dirichlet distribution, DP does not require pre-specifying number of components
- we’ll now describe how a DP generates a distribution over an unbounded set of items

# Running Example

- let's say we're trying to segment words into morphological units without any supervision:
  - walking → walk + ing
  - restarted → re + start + ed
- what is the unbounded set of latent items here?
  - lexicon of possible morphological units

# Dirichlet Process (DP)

- contains a “base distribution”  $G_0$
- simple example base distribution for our morph lexicon:

$$G_0(m) = p_{\text{len}}(|m|) \prod_{i=1}^{|m|} p_{\text{char}}(m_i)$$

- e.g., probability of “ing”:

$$G_0(\text{“ing”}) = p_{\text{len}}(3) p_{\text{char}}(\mathbf{i}) p_{\text{char}}(\mathbf{n}) p_{\text{char}}(\mathbf{g})$$

# Dirichlet Processes

- our unbounded distribution over items will choose its items by sampling from the base distribution
- base distribution typically has an infinite set of items with nonzero probability, as in our example:

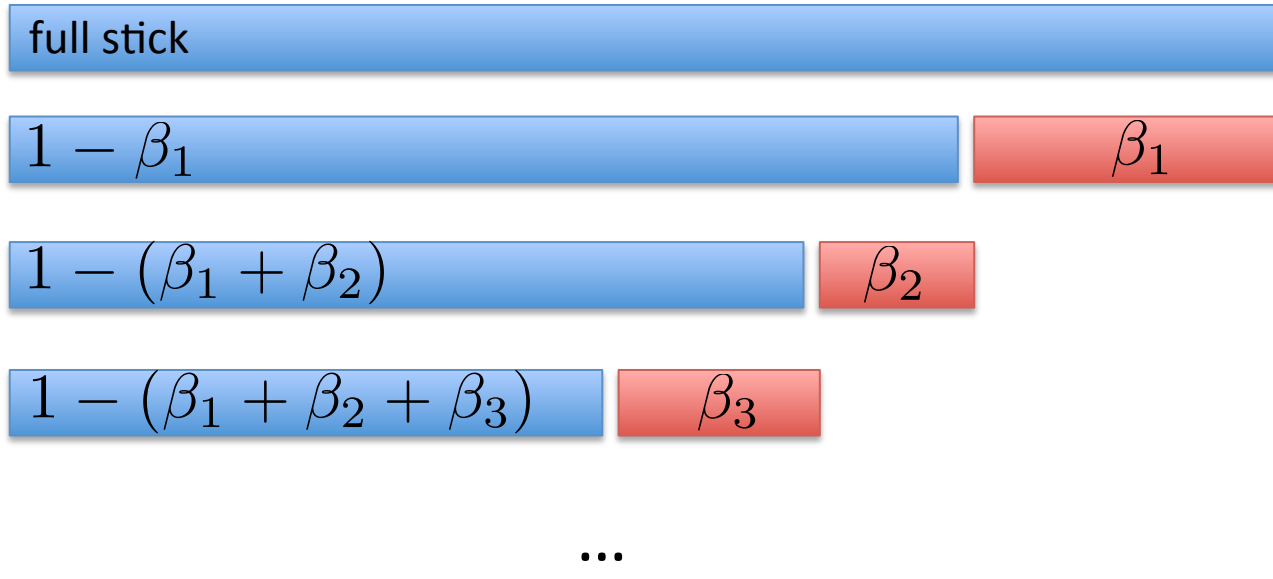
$$G_0(m) = p_{\text{len}}(|m|) \prod_{i=1}^{|m|} p_{\text{char}}(m_i)$$



# Items and Probabilities

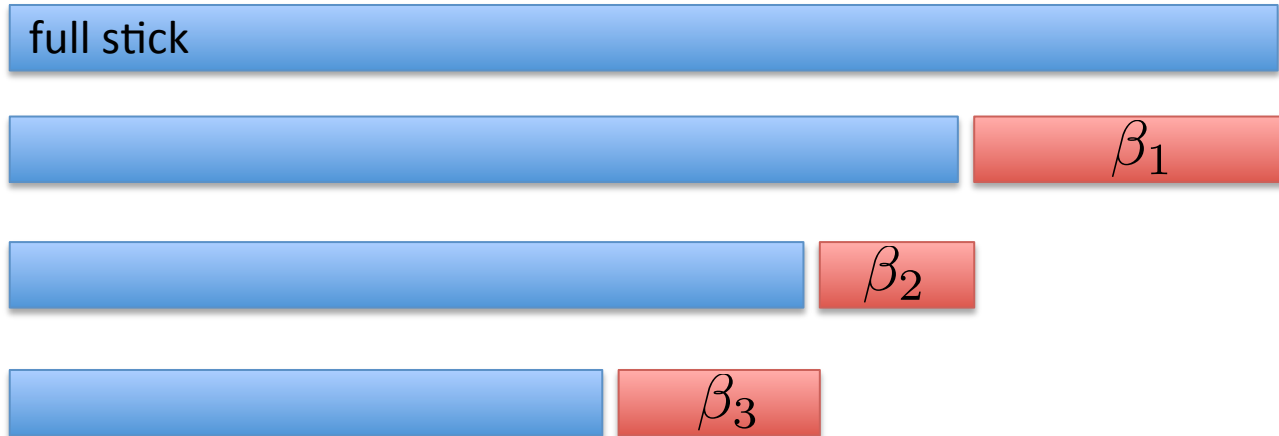
- base distribution provides the items (“atoms”), as many as we want
- where do their probabilities come from?
- we need an infinite set of probabilities that sum to 1
- DPs have another parameter: concentration (strength) parameter  $s$

# Stick-Breaking Process



- the betas form an infinite sequence that sums to 1
- they provide probabilities for an infinite set of items!

# Stick-Breaking Process



$$\nu_k \sim \text{Beta}(1, s)$$

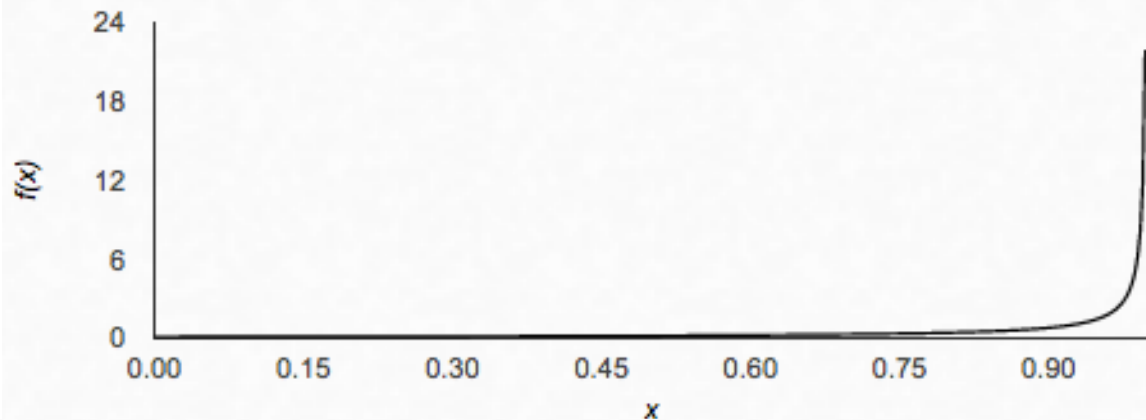
$$\beta_k = \nu_k \prod_{j=1}^{k-1} (1 - \nu_j)$$

- the betas form an infinite sequence that sums to 1
- they provide probabilities for an infinite set of items!

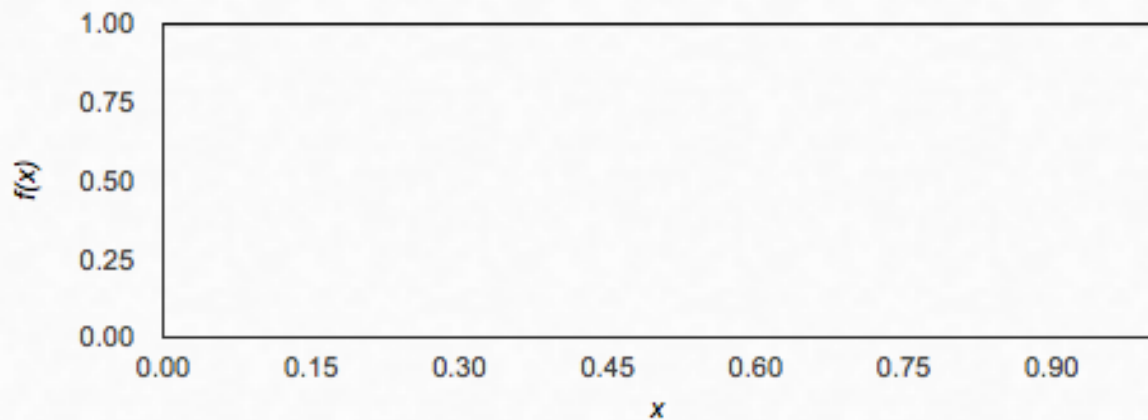
$$\nu_k \sim \text{Beta}(1, s)$$

## Beta Distribution

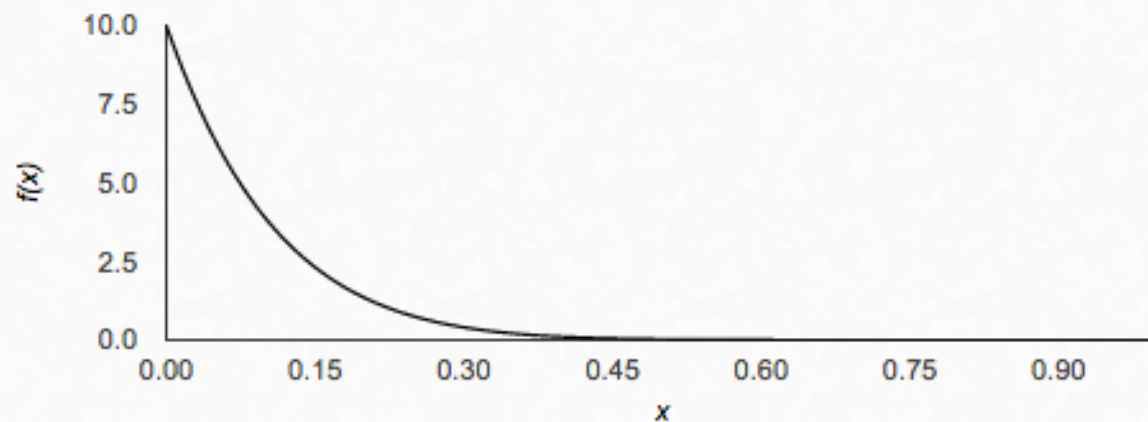
$$x \sim \text{Beta}(1, 0.1)$$



$$x \sim \text{Beta}(1, 1)$$



$$x \sim \text{Beta}(1, 10)$$



# Stick-Breaking with High Concentration ( $s = 10$ )

full stick

$$\nu_k \sim \text{Beta}(1, s)$$

$$\beta_k = \nu_k \prod_{j=1}^{k-1} (1 - \nu_j)$$

# Stick-Breaking with High Concentration ( $s = 10$ )



- high concentration = more of probability mass preserved for other pieces in the stick

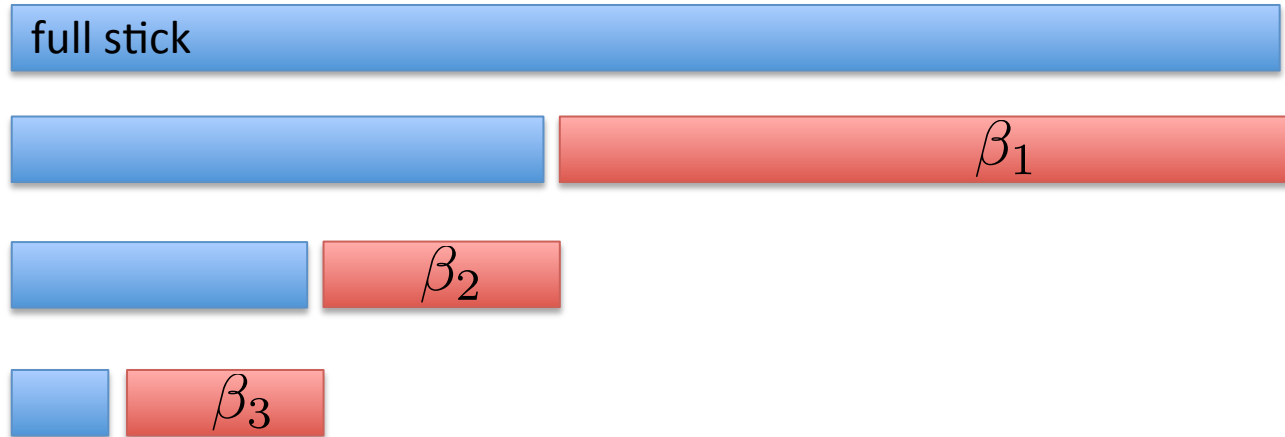
# Stick-Breaking with Low Concentration ( $s = 0.1$ )

full stick

$$\nu_k \sim \text{Beta}(1, s)$$

$$\beta_k = \nu_k \prod_{j=1}^{k-1} (1 - \nu_j)$$



# Stick-Breaking with Low Concentration ( $s = 0.1$ )



- low concentration = stronger power law effects in resulting probabilities





# A Draw $G$ from a DP

- 1:  $\beta \sim \text{GEM}(s)$   draw infinite probabilities from stick-breaking process with parameter  $s$
- 2:  $\theta_1, \theta_2, \dots \sim G_0$   draw atoms from base distribution  
atoms can be repeated!
- 3: the distribution  $G$  is defined as:

$$G(\theta) = \sum_{k=1}^{\infty} \beta_k \mathbb{I}[\theta = \theta_k]$$

# A Draw $G$ from a DP


- 1:  $\beta \sim \text{GEM}(s)$   draw infinite probabilities from stick-breaking process with parameter  $s$
- 2:  $\theta_1, \theta_2, \dots \sim G_0$   draw atoms from base distribution  
atoms can be repeated!
- 3: the distribution  $G$  is defined as:

$$G(\theta) = \sum_{k=1}^{\infty} \beta_k \mathbb{I}[\theta = \theta_k]$$

$$G(\text{"ing"}) = \sum_{k=1}^{\infty} \beta_k \mathbb{I}[\text{"ing"} = \theta_k]$$

- the stick-breaking construction of the DP is useful for specifying models and defining inference algorithms

# Dirichlet Process Mixture Model

- generative story for dataset  $x^{(1)}, x^{(2)}, \dots, x^{(n)}$  :
  - 1:  $\beta \sim \text{GEM}(s)$
  - 2:  $\theta_1, \theta_2, \dots \sim G_0$   what should the base distribution be?
  - 3: for  $i = 1 \dots n$ ,  $z^{(i)} \sim \beta$
  - 4: for  $i = 1 \dots n$ ,  $x^{(i)} \sim p(x^{(i)} \mid \theta_{z^{(i)}})$
- each  $x$  is generated from a single mixture component
- the number of mixture components is unbounded