A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes

Author: Yee Whye Teh, 2006

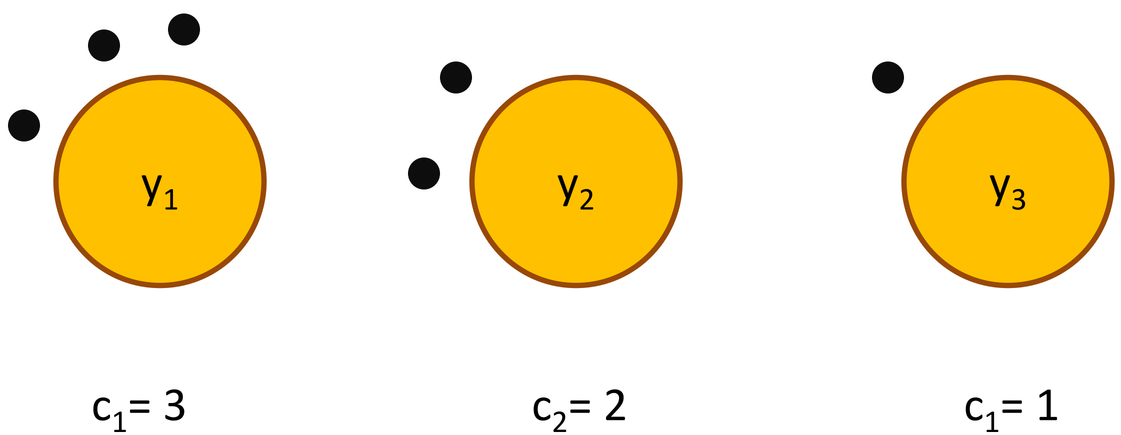
Reviewer: Xueqing Liu

Dirichlet Process (CRP) Recap

Model sequence of words as a sequence of customers coming to a restaurant: x_1, x_2, ……

Model vocabulary set as a sequence of tables (dishes) : y_1, y_2, ……

The $c. = \sum_{k=1}^{t} c_k$th customer comes and chooses table according to customer number in each table: $\frac{c_k}{\alpha + c.}$, chose a new table: $\frac{\alpha}{\alpha + c.}$



Pitman-Yor Process: a generalization

$$G \sim PY(d, \alpha, G_0)$$

G: customers sequence; G_0: tables sequence

Probability customer sits at table y_k: $\frac{c_k - d}{\alpha + c.}$; chooses new table: $\frac{\alpha + dt}{\alpha + c.}$

Assumes a **finite** vocabulary set/table sequence of size V: W

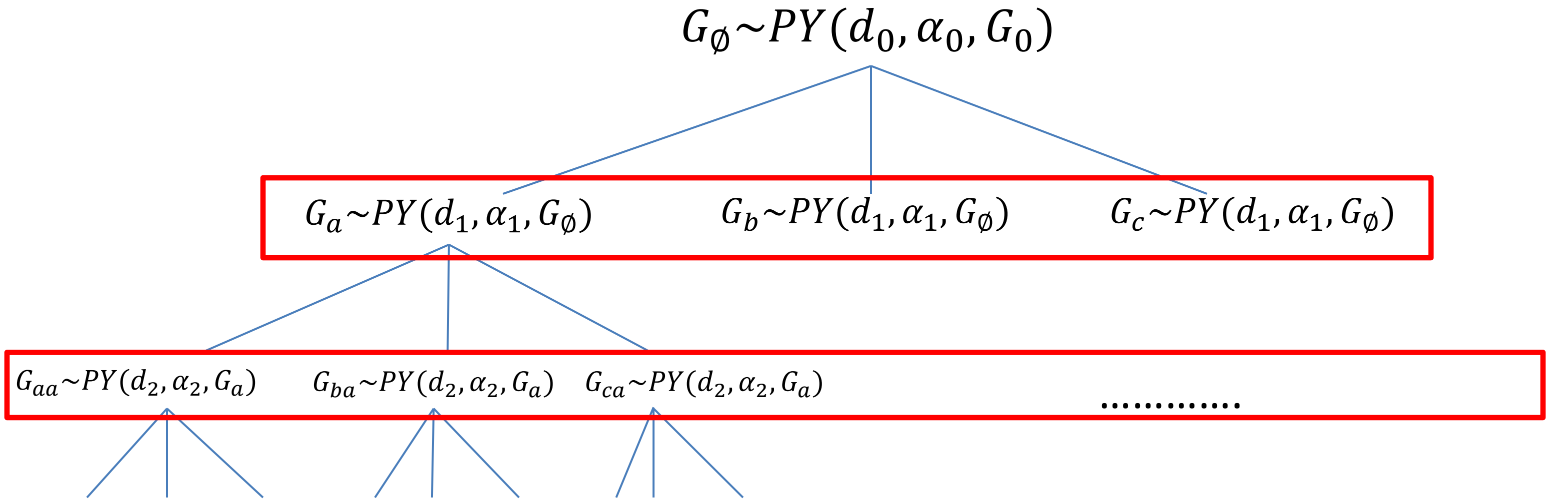
Power-law distribution: "rich-gets-richer"; number of unique words scale exponentially as $O(\alpha T^d)$

Hierarchical Pitman-Yor Language Model

$$G_{\boldsymbol{u}} \sim PY(d_{|\boldsymbol{u}|}, \alpha_{|\boldsymbol{u}|}, G_{\pi(\boldsymbol{u})})$$

Draw a sequence of customers $G_{\boldsymbol{u}}$ from another sequence of customers $G_{\pi(\boldsymbol{u})}$, $\pi(\boldsymbol{u}) = \boldsymbol{u}_2\boldsymbol{u}_3 \ldots \boldsymbol{u}_{|\boldsymbol{u}|}$

Consider W = {a, b, c}



Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}

G_{cac}	$x_{\mathbf{u}1}$	$x_{\mathbf{u}2}$	$x_{\mathbf{u}3}$	…….
	?	?	?	

Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}

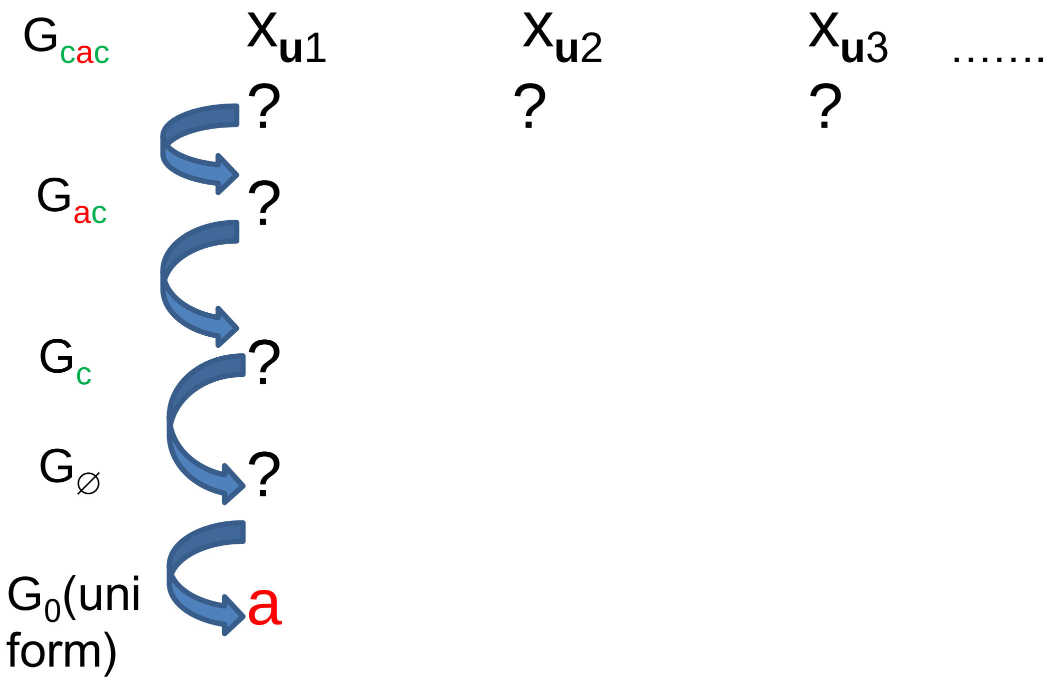
G_{cac}	$x_{\mathbf{u}1}$	$x_{\mathbf{u}2}$	$x_{\mathbf{u}3}$	…….
	?	?	?	
G_{ac}	?			

Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}



Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}

G_{cac}	$x_{\mathbf{u}1}$ =a	$x_{\mathbf{u}2}$?	$x_{\mathbf{u}3}$?	…….
G_{ac}	a	?		
G_{c}	a			
$G_{\varnothing}$	a			
G_0(uniform)	a			

$\frac{1-d_3}{\alpha_3+1}$

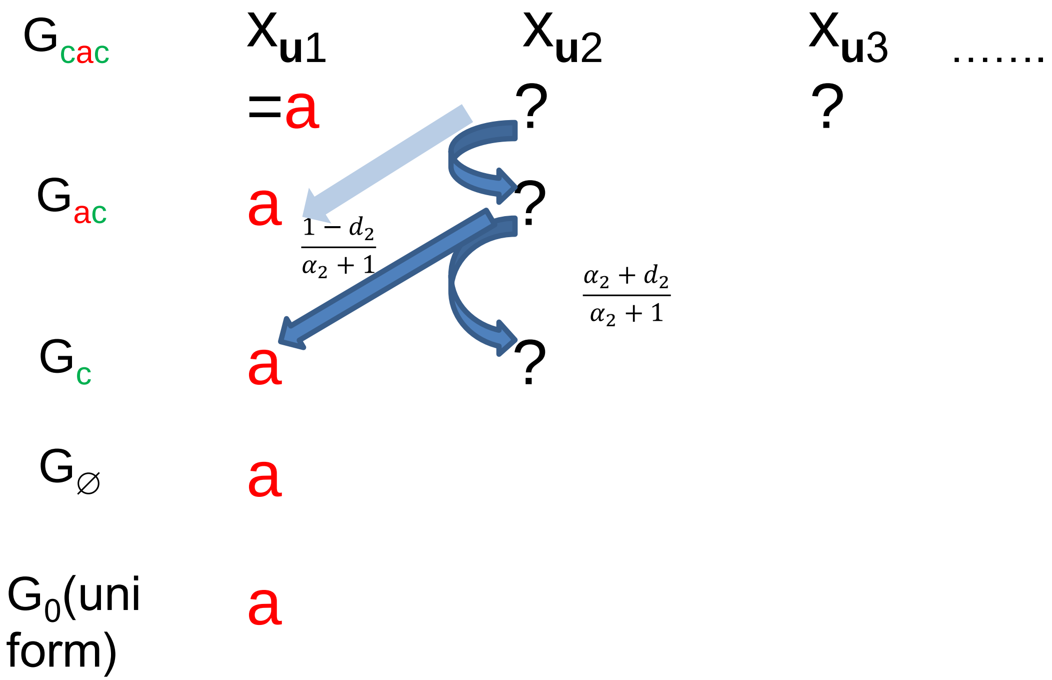
$\frac{\alpha_3+d_3}{\alpha_3+1}$

Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}

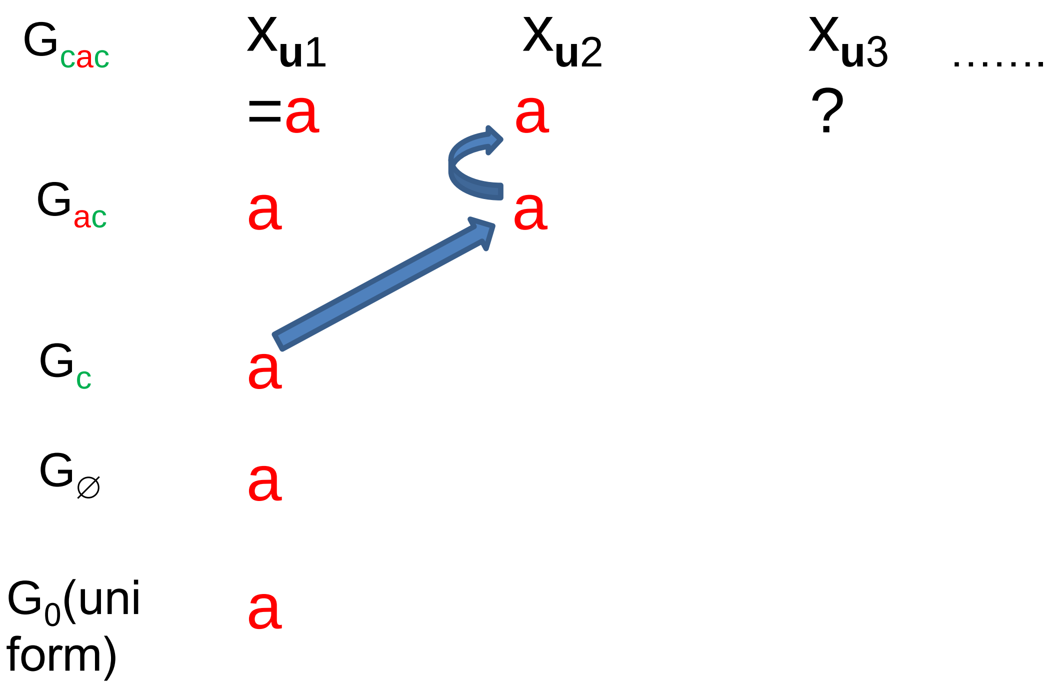


Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}

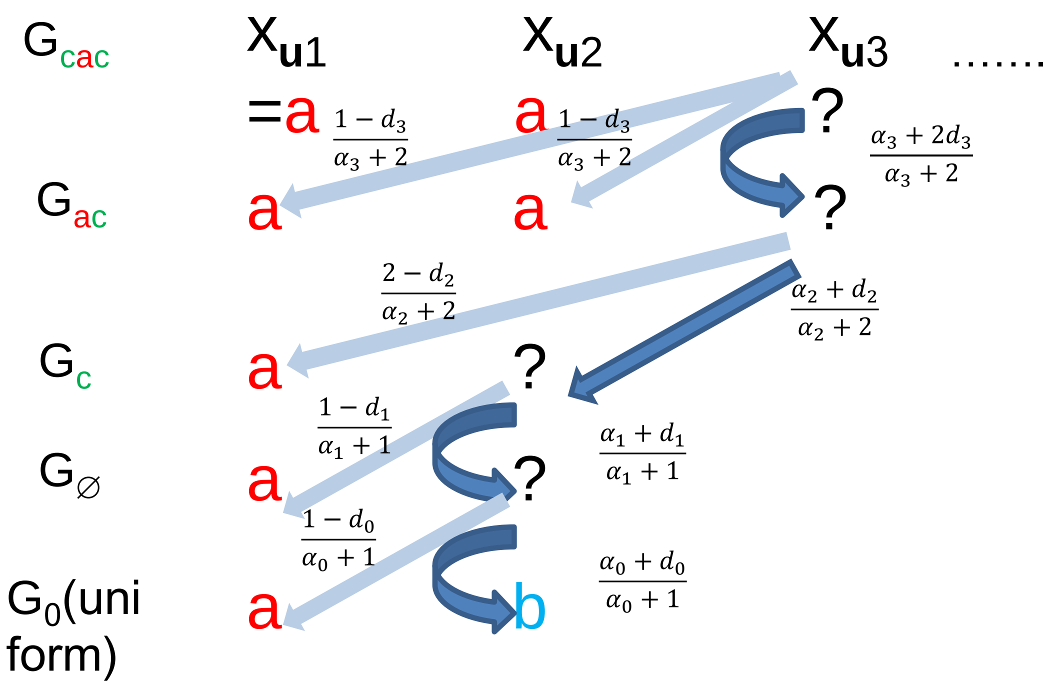


Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}



Hierarchical CRP: an Example

W = { a, b, c }

Context **u** = c a c ?

Sequence $x_{\mathbf{u}1}$, $x_{\mathbf{u}2}$, ……drawn from G_{cac}

G_{cac}	$x_{\mathbf{u}1}$ =a	$x_{\mathbf{u}2}$ a	$x_{\mathbf{u}3}$ b	…….
G_{ac}	a	a	b	
G_{c}	a	b		
$G_{\varnothing}$	a	b		
G_0(uniform)	a	b		

Inference with Gibbs Sampling

G_us are marginalized out; use S_u, $\boldsymbol{S} = \{S_v\}$, $\boldsymbol{\Theta} = \{\alpha_m, d_m\}$

$$p(w|\boldsymbol{u}, D) = \int p(w|\boldsymbol{u}, \boldsymbol{S}, \boldsymbol{\Theta})\, p(\boldsymbol{S}, \boldsymbol{\Theta}|D) d(\boldsymbol{S}, \boldsymbol{\Theta})$$

Approximate the integral with samples:

$$p(w|\boldsymbol{u}, D) \approx \sum_i p\left(w\middle|\boldsymbol{u}, \boldsymbol{S}^{(i)}, \boldsymbol{\Theta}^{(i)}\right)$$

Recursively compute $p(w|\boldsymbol{u}, \boldsymbol{S}, \boldsymbol{\Theta})$:

$$p(w|\boldsymbol{u}, \boldsymbol{S}, \boldsymbol{\Theta}) = \frac{c_{\boldsymbol{u}w\cdot} - d_{|\boldsymbol{u}|} t_{\boldsymbol{u}w\cdot}}{\theta_{|\boldsymbol{u}|} + c_{\boldsymbol{u}\cdot\cdot}} + \frac{\theta_{|\boldsymbol{u}|} + d_{|\boldsymbol{u}|} t_{\boldsymbol{u}w\cdot}}{\theta_{|\boldsymbol{u}|} + c_{\boldsymbol{u}\cdot\cdot}} p(w|\pi(\boldsymbol{u}), \boldsymbol{S}, \boldsymbol{\Theta})$$

Inference with Gibbs Sampling

Gibbs sampling:

$$p(k_{\boldsymbol{u}l} = k \mid \boldsymbol{S}^{-\boldsymbol{u}l}, \boldsymbol{\Theta}) \propto \frac{\max\left(0, c_{\boldsymbol{u}x_{\boldsymbol{u}l}k}^{-\boldsymbol{u}l} - d\right)}{\theta + c_{\boldsymbol{u}\cdot\cdot}^{-\boldsymbol{u}l}}$$

$$p\left(k_{\boldsymbol{u}l} = k^{new} with\, y_{\boldsymbol{u}k^{new}} = x_{\boldsymbol{u}l} \middle| S^{-\boldsymbol{u}l}, \Theta\right) \propto \frac{\theta + dt_{u\cdot\cdot}^{-\boldsymbol{u}l}}{\theta + c_{u\cdot\cdot}^{-\boldsymbol{u}l}} p(x_{\boldsymbol{u}l} | \pi(\boldsymbol{u}), \boldsymbol{S}^{-\boldsymbol{u}l}, \boldsymbol{\Theta})$$

Experimental Results

IKN: interpolated Kneser-Ney

MKN: modified Kneser-Ney

HPYLM: Pitman-Yor using Gibbs sampler

HPYCV: parameters obtained by cross-validaion

T	n	IKN	MKN	HPYLM	HPYCV	HDLM
2e6	3	148.8	**144.1**	145.7	144.3	191.2
4e6	3	137.1	**132.7**	134.3	**132.7**	172.7
6e6	3	130.6	126.7	127.9	**126.4**	162.3
8e6	3	125.9	122.3	123.2	**121.9**	154.7
10e6	3	122.0	118.6	119.4	**118.2**	148.7
12e6	3	119.0	115.8	116.5	**115.4**	144.0
14e6	3	116.7	113.6	114.3	**113.2**	140.5
14e6	2	169.9	**169.2**	169.6	169.3	180.6
14e6	4	106.1	102.4	103.8	**101.9**	136.6