# A Hierarchical Bayesian Language Model based on Pitman-Yor Processes

Yee Whye Teh

*Presented by Hsin-Ta Wu*

*Slides courtesy: Yee Whye Teh*

# Language Model

- Given a sentence of $t$ words:

$$word_1, word_2, \dots, word_t$$

- An *n*-gram ***LANGUAGE MODEL*** defines a probability distribution over the current $word_i$ given the prior $n\text{-}1$ words.

$$P(word_i | word_{i-n+1}, \dots, word_{i-1})$$

- This sentence then can be typically represented by the probability:

$$P(word_1, word_2, \dots, word_t) = \prod_{i=1}^{t} P(word_i | word_{i-n+1}, \dots, word_{i-1})$$

# Language Model (cont)

- Consider a set vocabulary $W$ with $V$ word types
- Each word $\mathbf{w} \in W$, and a context $\mathbf{u}$:*n-1* prior-word
  - **E.g. n=3, <u>bayesian nonparametric</u> <u>model</u>**

    $\underbrace{\text{bayesian nonparametric}}_{\mathbf{u}}\ \underbrace{\text{model}}_{\mathbf{w}}$
- The vector of word probability estimates for n-grams:

$$G_u = [G_u(w)]_{w \in W} = [G_u(w_1), \ldots, G_u(w_v)]$$

- Maximum Likelihood estimation:

$$P(word_i = w | word_{i-n+1}, \ldots, word_{i-1} = u)$$

$$= G_u^{ML}(w) = \frac{c_{\boldsymbol{u}w}}{\sum_{W'} c_{\boldsymbol{u}w'}} = \frac{c_{\boldsymbol{u}w}}{c_{\boldsymbol{u}\cdot}}$$

# Smoothing

- Maximum Likelihood is expected to be a very poor estimate given a realistic corpus size
  - What about a trigram $uw$ which has never occurred in the training data
    - i.e. $G_u^{ML}(w) = 0$
- *Smoothing* is used to address this problem.

$$G_u^{ML}(w) = \frac{\delta + c_{\boldsymbol{u}w}}{\delta|V| + c_{\boldsymbol{u}\cdot}}$$

# Back-off and Interpolated Smoothing

- Back-off approach:
  - Only use lower-order model when data for higher- order model is unavailable (i.e. count is zero).

$$P_{katz}(w_n \mid w_{n-N+1}^{n-1}) = \begin{cases} P^*(w_n \mid w_{n-N+1}^{n-1}) & \text{if } C(w_{n-N+1}^{n-1}) > 1 \\ \alpha(w_{n-N+1}^{n-1}) P_{katz}(w_n \mid w_{n-N+2}^{n-1}) & \text{otherwise} \end{cases}$$

- Interpolated approach:
  - Linearly combine estimates of *n*-gram models of increasing order.

$$\hat{P}(w_n \mid w_{n-2}, w_{n-1}) = \lambda_1 P(w_n \mid w_{n-2}, w_{n-1}) + \lambda_2 P(w_n \mid w_{n-1}) + \lambda_3 P(w_n)$$

$$\text{Where: } \sum_i \lambda_i = 1$$

# Bayesian Smoothing

- Estimation

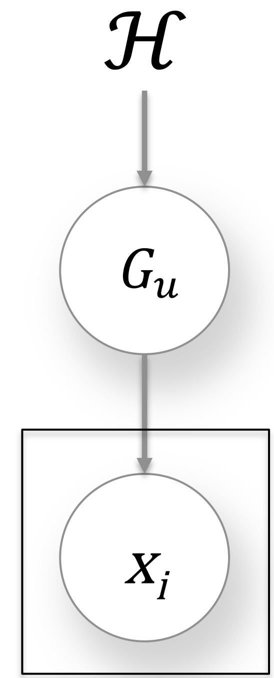
$$P(G_u|\mathcal{D}) \propto P(\mathcal{D}|G_u)P(G_u)$$

- Predictive Inference

$$P(word_i = w|word_{i-n+1}, \ldots, word_{i-1} = u, \mathcal{D})$$
$$= \int P(w|u, G_u)P(G_u|\mathcal{D})dG_u$$

- Priors over distributions

$$G_u \sim \mathcal{DP}(\theta, \mathcal{H})$$
$$G_u \sim \mathcal{PY}(d, \theta, \mathcal{H})$$




- Inference is smoothed with respect to the distribution

# Pitman-Yor Process

- *Pitman-Yor Process*

  $\mathcal{PY}(d, \theta, G_0)$

  - $d$: discount parameter, $0 \leq d < 1$
  - $\theta$: strength (concentration) parameter, $\theta > -d$
  - $G_0$: base distribution

- Generalization of the *Dirichlet process* $(\mathrm{d}=0)$

- Pitman-Yor processes produce distributions over words given by a power law distribution
  - [Goldwater et al 2006] investigated the Pitman-Yor process from this perspective.

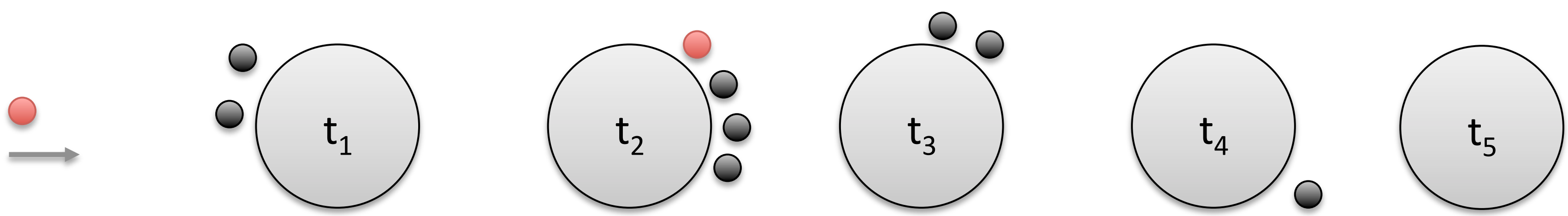
# Pitman-Yor Process for **a unigram language model**

- To estimate a word $\mathbf{w} \in W$,
  - $P(word_i = w | word_{i-n+1}, \ldots, word_{i-1} = u)$
    $= P(word_i = w) = G(w)$
  - $G = [G(w)]_{w \in W}$

- $G \sim \mathcal{PY}(d, \theta, G_0)$
  - $d$: discount parameter, $0 \leq d < 1$
  - $\theta$: strength parameter, $\theta > -d$
  - $G_0$: a mean vector for unigram, using uniform distribution over fixed vocabulary W of V words

# Perspective by the Chinese restaurant process

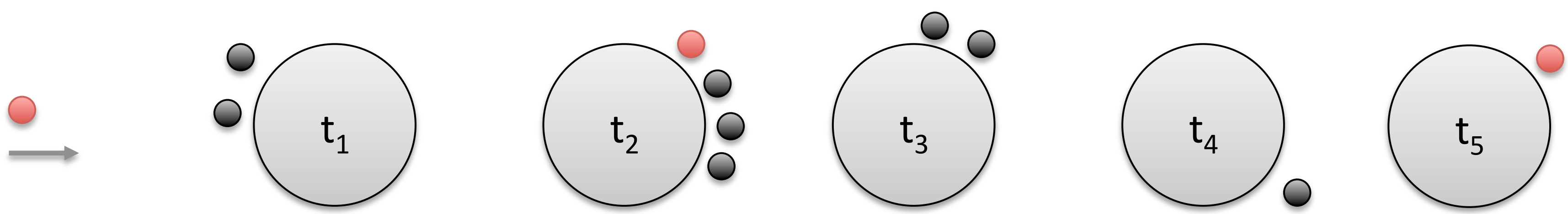
- Easiest to understand them using Chinese restaurant processes.




$$P(sit\ at\ an\ occupied\ table\ t_i) = \frac{c_{t_i} - d}{\theta + c_.}$$

# Perspective by the Chinese restaurant process

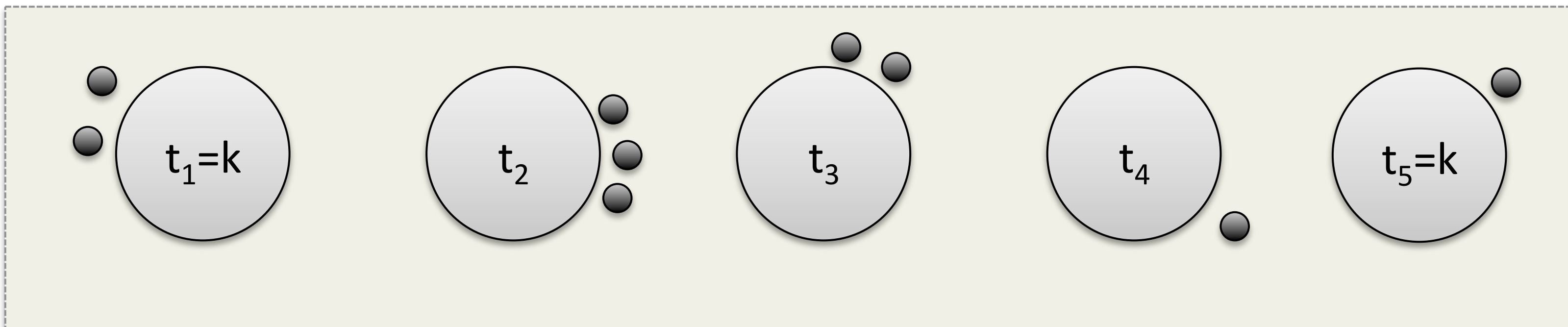
- Easiest to understand them using Chinese restaurant processes.




$$P(sit\ at\ an\ occupied\ table\ t_i) = \frac{c_{t_i} - d}{\theta + c_.} \qquad P(sit\ at\ new\ table) = \frac{\theta + dt_.}{\theta + c_.}$$

# Perspective by the Chinese restaurant process

- Given the seating arrangement ***S***, the predictive probability of a test word ***k*** is:




$$P(x_{c.+1} = k|S) = \frac{c_k - dt_k}{\theta + c.} + \frac{\theta + dt.}{\theta + c.} G_0(k)$$

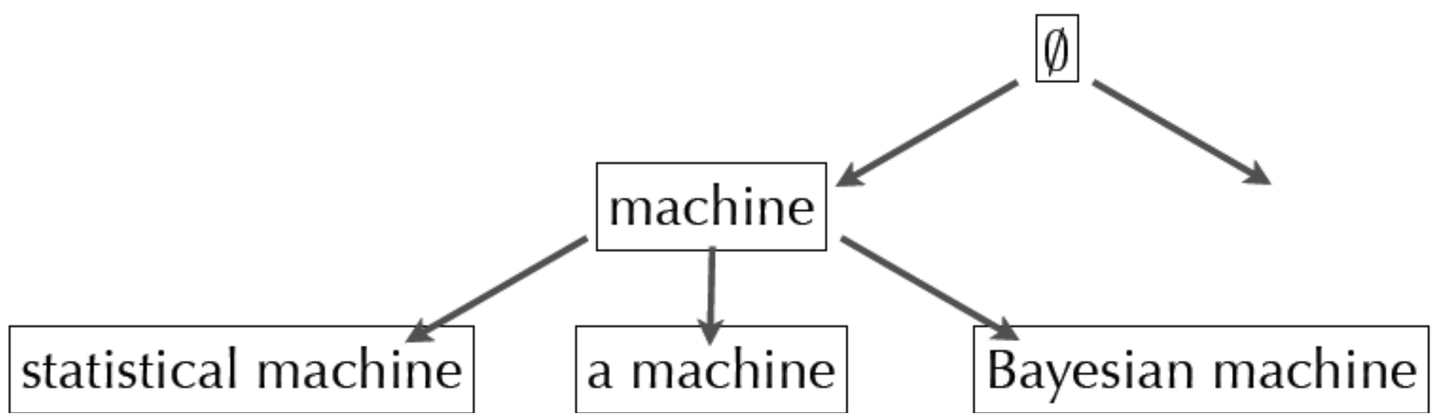
# What about *n*-gram language model?

- Hierarchical Bayesian models
  - Capture the dependencies by statistical strength among different components of the language model
  - Specifically: hierarchical model based on the tree of suffixes : CONTEXT TREES

# Context Tree

- Basic assumption: words appearing later in a context are more important

*Xxxxxxx x xxxxxxx in statistical machine learning. Xxxxxx xx x x xx. Bayesian machine xxx xx. Xxxx Is a machine….*



# Hierarchical Bayesian Models on Context Tree

**[MacKay and Peto 1994]**

- The probability of the current word *w* following the context *u*

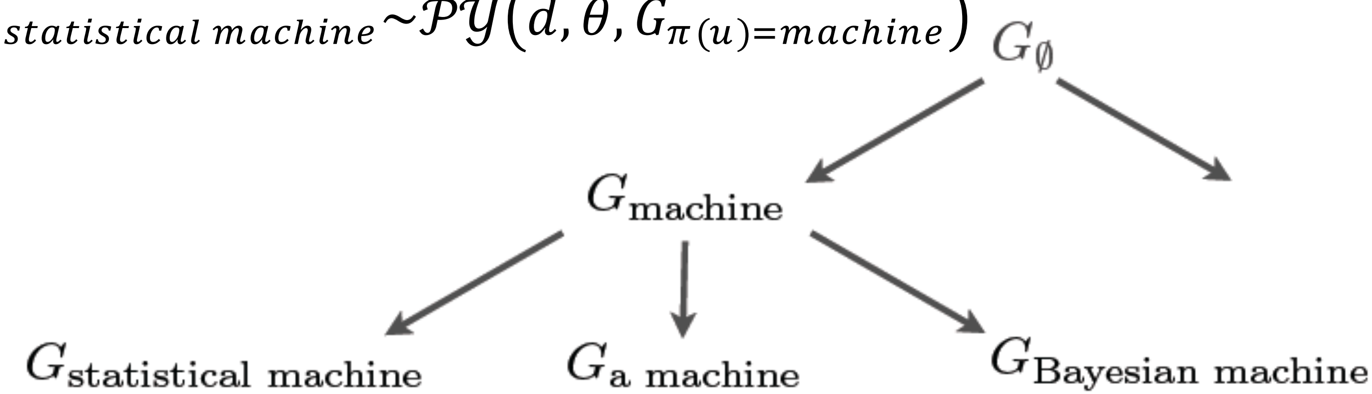
$$P(word_i = w | word_{i-n+1}, \dots, word_{i-1} = u) = G_u(w)$$

- The vector of word probability estimates for *n*-grams

$$G_u = [G_u(w)]_{w \in W} = [G_u(w_1), \dots, G_u(w_v)]$$

- Tie related distribution together

$$G_{u=statistical\ machine} \sim \mathcal{DP}\left(\theta, G_{\pi(u)=machine}\right)$$
$$G_{u=statistical\ machine} \sim \mathcal{PY}\left(d, \theta, G_{\pi(u)=machine}\right)$$



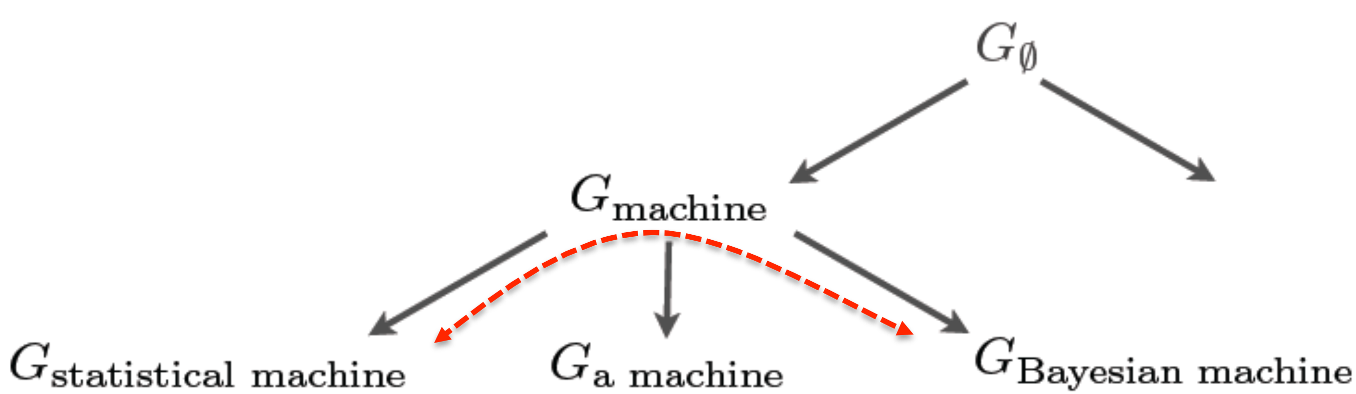
# Hierarchical Bayesian Models on Context Tree

**[MacKay and Peto 1994]**

- Tie related distribution together

$$G_{statistical\ machine} \sim \mathcal{DP}(\theta, G_{machine})$$
$$G_{statistical\ machine} \sim \mathcal{PY}(d, \theta, G_{machine})$$

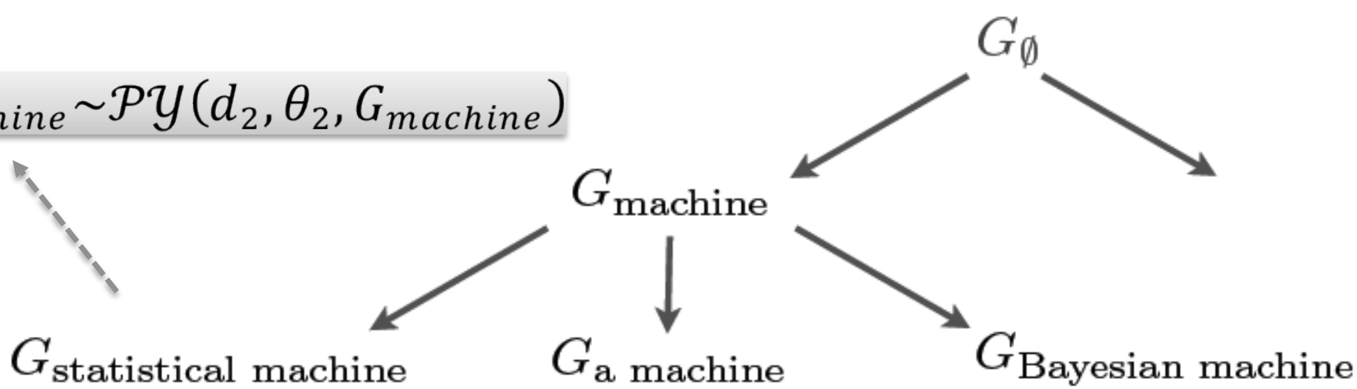
  - Observations in one context affect inference in other context.
  - Statistical strength is shared between similar contexts
  - E.g. Observe "statistical machine learning"



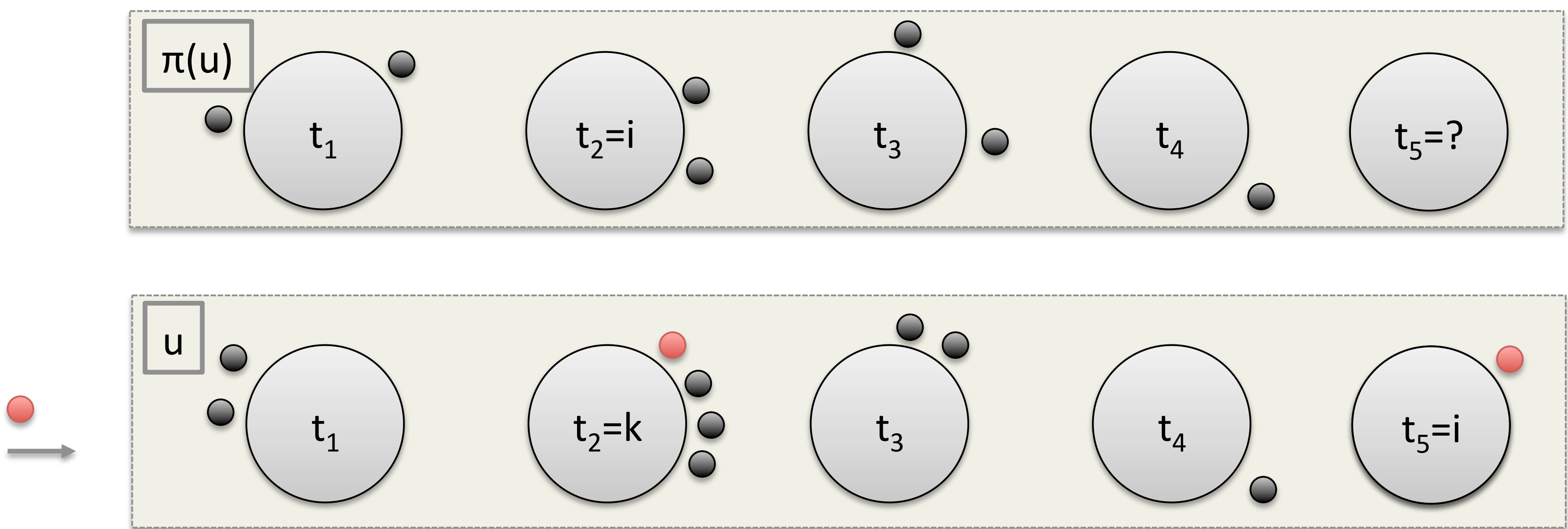
# Hierarchical Pitman-Yor Process for ***n*-gram Language Models**

- Use a Pitman-Yor process as the prior for each node $G_u = [G_u(w)]_{w \in W}$
- $G_u \sim \mathcal{PY}\left(d_{|u|}, \theta_{|u|}, G_{\pi(u)}\right)$

$G_{statistical\ machine} \sim \mathcal{PY}(d_2, \theta_2, G_{machine})$



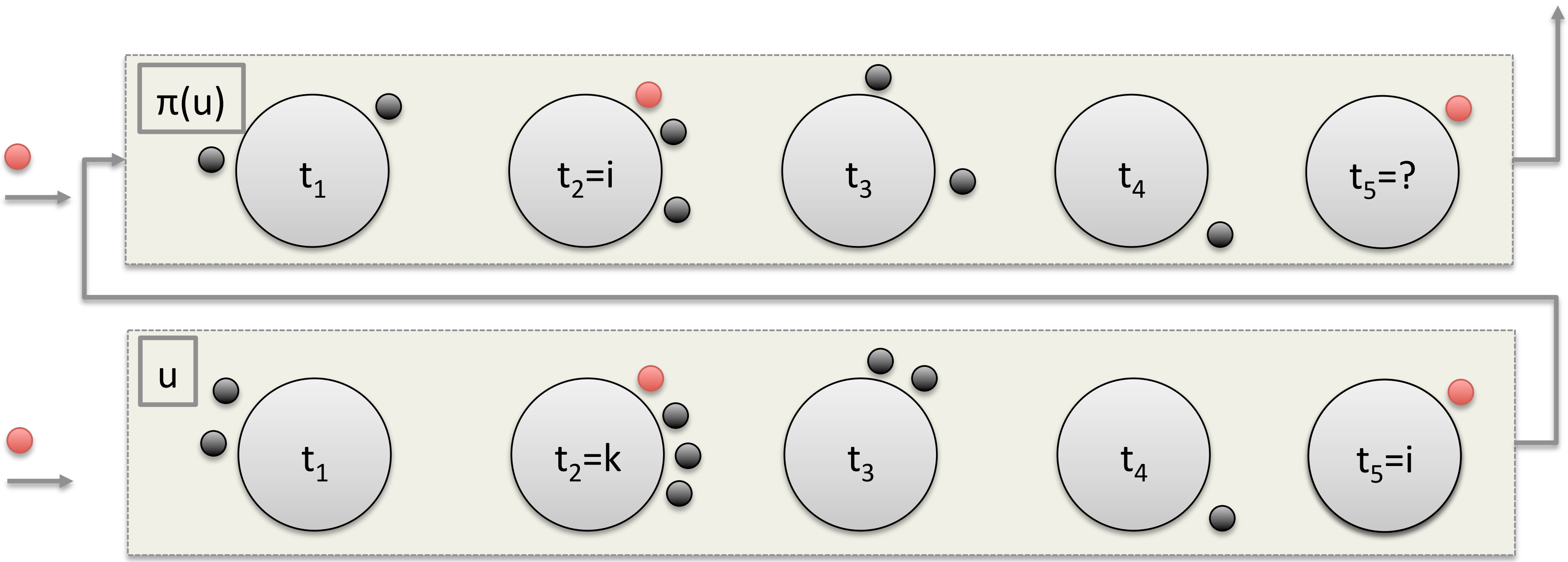
# Perspective by the Chinese restaurant process




$$P(sit\ at\ an\ occupied\ table\ k) = \frac{c_{uwk} - d_{|u|}}{\theta_{|u|} + c_{u\cdot\cdot}}$$

$$P(sit\ at\ a\ new\ table) = \frac{\theta_{|u|} + d_{|u|} t_{u\cdot}}{\theta_{|u|} + c_{u\cdot\cdot}}$$

# Perspective by the Chinese restaurant process




$$P(sit\ at\ an\ occupied\ table\ i) = \frac{c_{\pi(u)wi} - d_{|\pi(u)|}}{\theta_{|\pi(u)|} + c_{\pi(u)\cdot\cdot}}$$

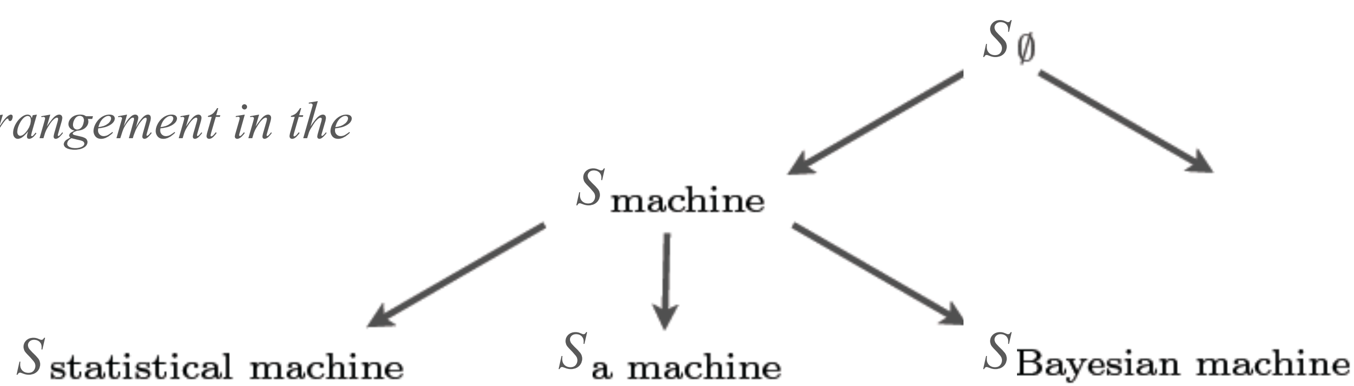
$$P(sit\ at\ a\ new\ table) = \frac{\theta_{|\pi(u)|} + d_{|\pi(u)|} t_{\pi(u)\cdot}}{\theta_{|\pi(u)|} + c_{\pi(u)\cdot\cdot}}$$

# Hierarchical Pitman-Yor Process for ***n*-gram Language Models**

- Given a particular seating arrangement,

$$P(w = learning \mid u = statistical\ machine)$$
$$= \frac{c_{uw\cdot} - d_{|u|} t_{uw\cdot}}{\theta_{|u|} + c_{u\cdot\cdot}} + \frac{\theta_{|u|} + d_{|u|} t_{u\cdot\cdot}}{\theta_{|u|} + c_{u\cdot\cdot}} P(w = learning \mid \pi(u) = machine)$$

$S_u$: *seating arrangement in the restaurant* ***u***



# What's next? Inference

- Based on the framework for Hierarchical Pitman-Yor Language Model, to get the probability over a word w after a context u P(*w*|*u*) given training data *D*:

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

  - inference of seating arrangements ***S*** in each restaurant
  - estimation of the context-specific parameters $\mathbf{\Theta}$

# Inference of Seating Arrangements

- Gibbs sampling is used to keep track of which table each customer sits at

- Steps:
  - Iterative over all customers present in each restaurant, resampling the table at which each customer sits
    - Randomly removing a customer from the restaurant
    - Then resampling the table at which that customer sits

# Estimation of the context parameters

- For a *n*-gram language model, there are *2n* parameters

$$\Theta = \{d_m, \theta_m : 0 \leq m \leq n-1\}$$

- Use the auxiliary variable sampling method, assuming $\theta_m \sim Gamma(\alpha_m, \beta_m)$ $d_m \sim Beta(a_m, b_m)$

- Further details please find the technical report [Teh, 2006]

# The predictive probability:

- Approximate the integral with samples $\{S^{(i)}, \Theta^{(i)}\}^{I}_{i=1}$ drawn from $p(S, \Theta/D)$:

$$p(w|\mathbf{u}, \mathcal{D}) \approx \sum_{i=1}^{I} p(w|\mathbf{u}, \mathcal{S}^{(i)}, \mathbf{\Theta}^{(i)})/I$$

# Interpolated Kneser-Ney (IKN) and Modified Kneser-Ney (MKN)

$$P_u^{ML}(w) = \frac{c_{uw}}{\sum_{w'} c_{uw'}} = \frac{c_{uw}}{c_{u\cdot}}$$

Modified Kneser-Ney (MKN)

$$D(c) = \begin{cases} 0 & \text{if } c = 0 \\ D_1 & \text{if } c = 1 \\ D_2 & \text{if } c = 2 \\ D_{3+} & \text{if } c \geq 3 \end{cases}$$

$$P_{\mathbf{u}}^{\text{IKN}}(w) = \frac{\max(0, c_{\mathbf{u}w} - d_{|\mathbf{u}|})}{c_{\mathbf{u}\cdot}} + \frac{d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{c_{\mathbf{u}\cdot}} P_{\pi(\mathbf{u})}^{\text{IKN}}(w)$$

$$P_{\mathbf{u}}^{\text{HPY}}(w \mid \text{seating arrangement}) = \frac{c_{\mathbf{u}w\cdot} - d_{|\mathbf{u}|} t_{\mathbf{u}w}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} + \frac{\theta_{|\mathbf{u}|} + d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} P_{\pi(\mathbf{u})}^{\text{HPY}}(w \mid \text{seating arrangement})$$

- Assume that the strength parameters $\theta_{|u|} = 0$ for all $\boldsymbol{u}$
- Restrict $t_{uw}$ to be at most 1
    - all customers representing the same word token should only sit on the same table in each restaurant

- Interpret IKN as an approximate inference scheme for the HPYLM

[Chen and Goodman. 1998. An empirical study of smoothing techniques for language modeling.

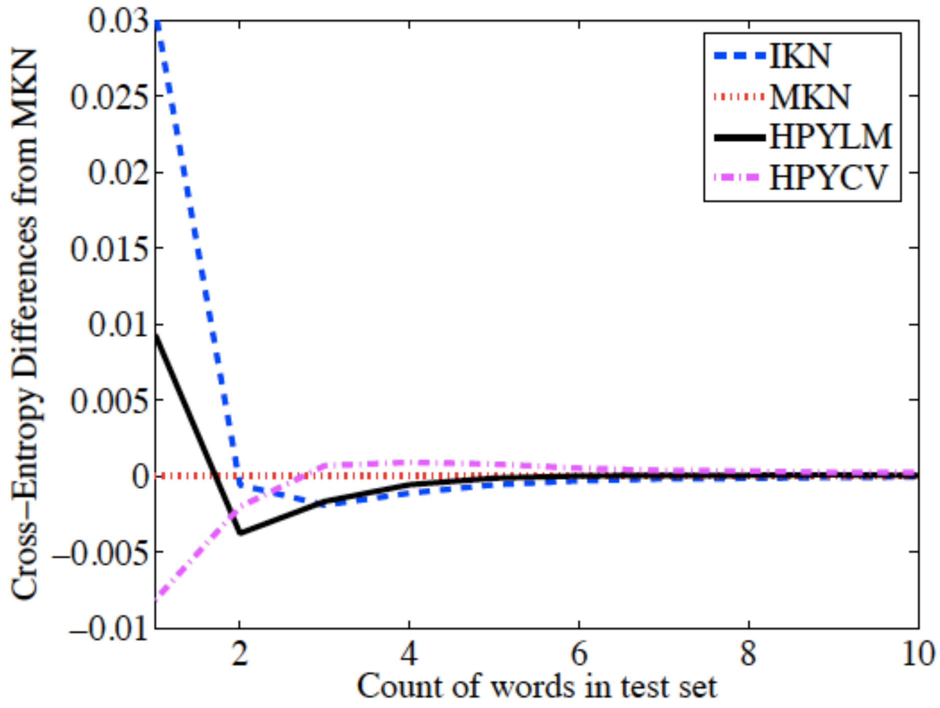
# Experiments

- Test five language models on APNews corpus:
  - Interpolated Kneser-Ney (IKN)
  - Modified Kneser-Ney (MKN)
  - Hierarchical Pitman-Yor Language Model (HPYLM)
  - Optimized HPYLM (HPYCV)
  - Hierarchical Dirichlet Language Model (HDLM)
- Evaluated by Perplexites
  - Train the *n*-Gram model: $p(w_i | w_{i-n+1}^{i-1})$
  - Calculate: $p(T) = \prod_{i}^{l_T} p(t_i)$
  - Cross-entropy: $H_p(T) = -\frac{1}{W_T} \log_2 p(T)$
  - Perplexity: $\mathrm{PP}_p(T) = 2^{H_p(T)}$

# Experimental Results I

| 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 |

# Experimental Results II

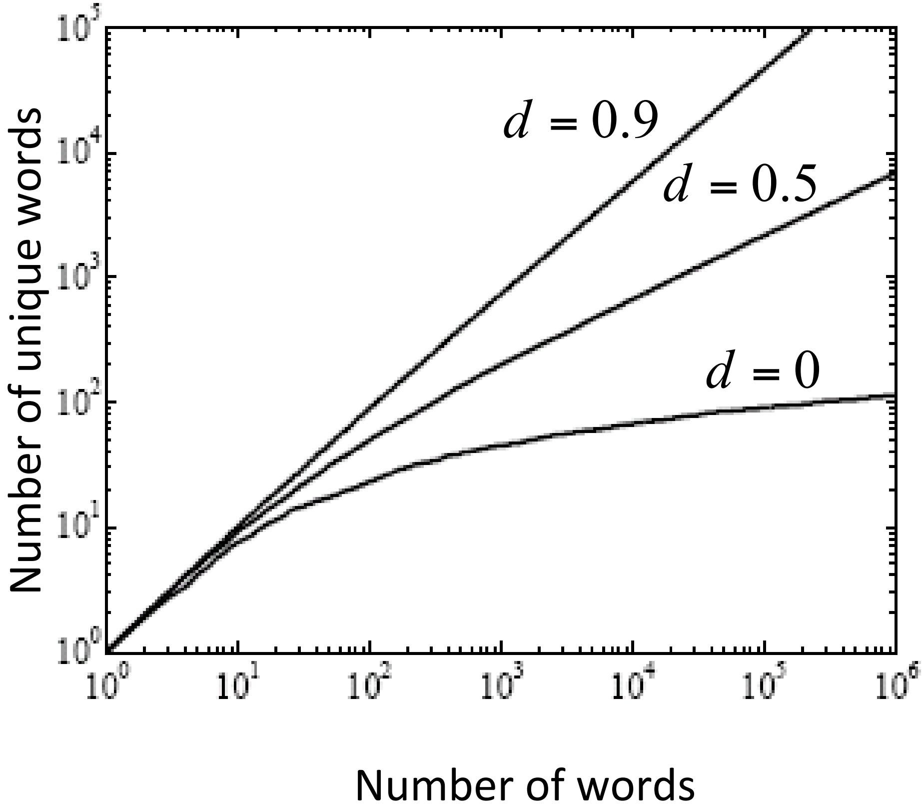


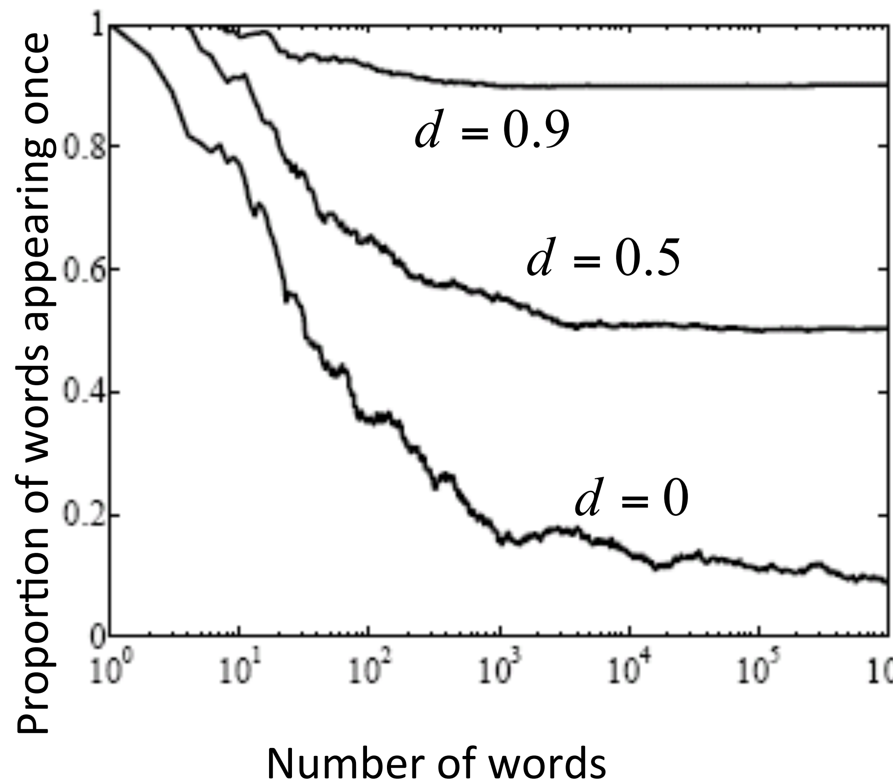
# Conclusions

- Proposed a new language model based on the hierarchical Bayesian paradigm.
- Showed that Interpolated Kneser-Ney is approximate inference in the hierarchical Pitman-Yor language model.

# QUESTIONS

# Power-law properties of the Pitman-Yor Process



# Experimental Results II