

Hierarchical Tree-Structured Stick-Breaking Priors

Hiroshi Noji^{1,3}

Daichi Mochihashi^{2,3}

Yusuke Miyao^{1,3}

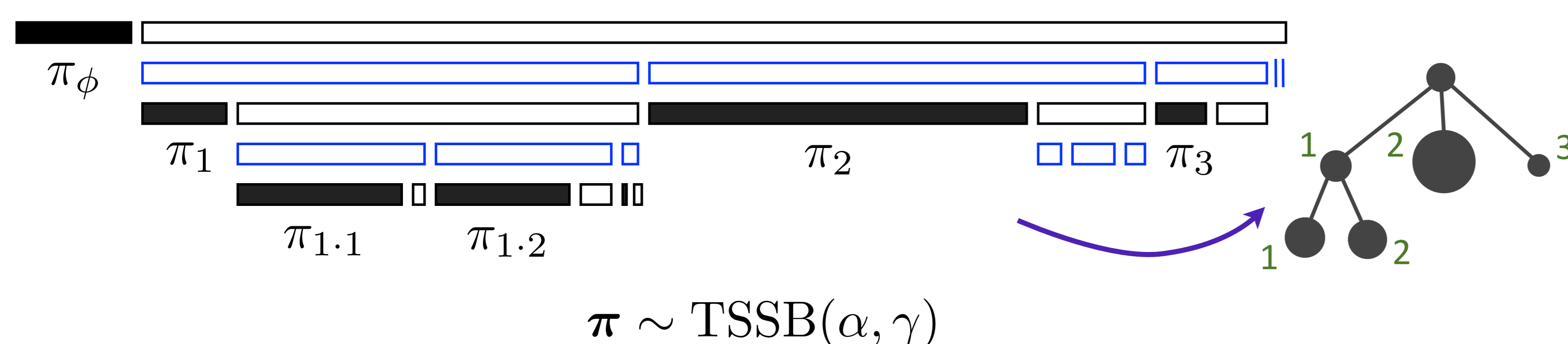
1. National Institute of Informatics, Tokyo
2. The Institute of Statistical Mathematics, Tokyo
3. Graduate University for Advanced Studies

Overview

- Current models ignore the relationships between hidden states
- e.g., the states of HMM or PCFG are exclusive
- Propose the general nonparametric prior which induces *the latent hierarchy between the hidden states*
- Construct a *HMM on a tree*, and a tree-structured topic model
- Topic model works, but the HMM currently fails \Rightarrow why?

Tree-Structured Stick-Breaking [1]

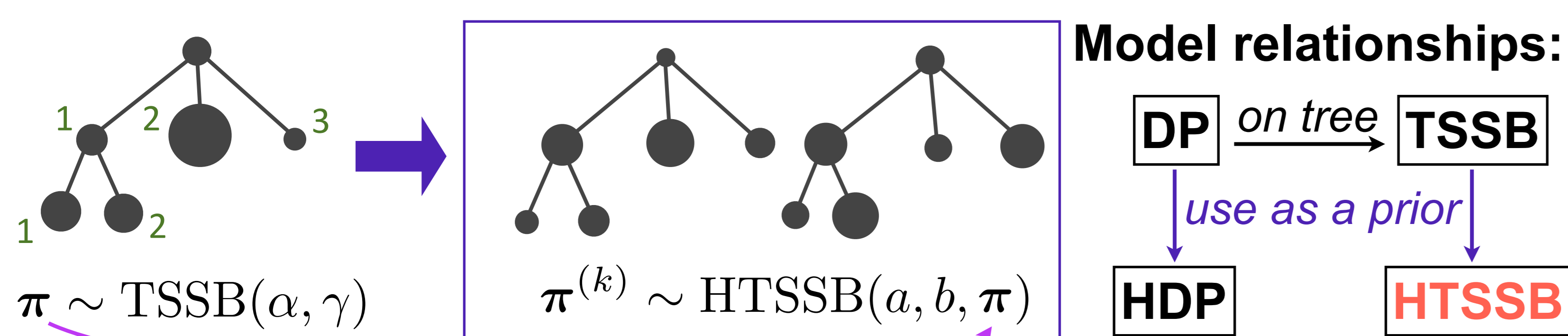
Partitions a unit interval hierarchically to get a measure on a tree



- The model consists of two kinds of stick-breakings:
 - ν -break selects stop or pass at the node: $\nu_\epsilon \sim \text{Beta}(1, \alpha)$
 - ψ -break selects the child direction: $\psi_\epsilon \sim \text{Beta}(1, \gamma)$
- Example: $\pi_{1.2} = (1 - \nu_\phi) \cdot \psi_1 \cdot (1 - \nu_1) \cdot (1 - \psi_{1.1}) \cdot \psi_{1.2} \cdot \nu_{1.2}$
- Generalization of the Dirichlet process on the tree
- **Problem:**
 - Each draws from this prior creates *a different tree structure*
 - The same problem when extending the Dirichlet process to the grouped data, e.g., HDP-HMM \Rightarrow *define another type of hierarchy!*

Hierarchical TSSB: core idea

Use a draw from the TSSB as a base measure of another draw

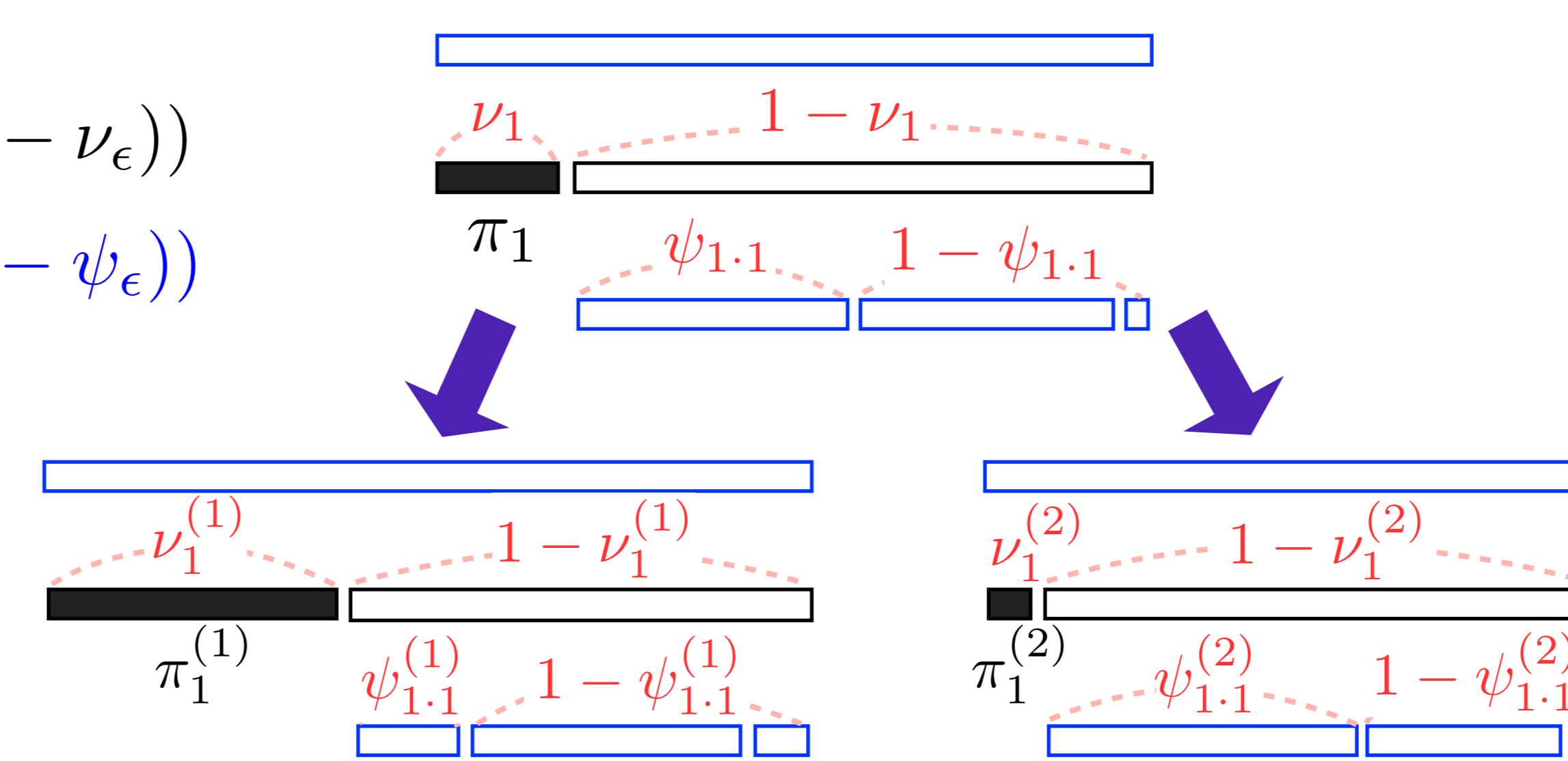


Hierarchical ν - and ψ -breaks

Stick lengths of base measure are used as a prior in each position

$$\nu_\epsilon^{(k)} \sim \text{Beta}(a\nu_\epsilon, a(1 - \nu_\epsilon))$$

$$\psi_\epsilon^{(k)} \sim \text{Beta}(b\psi_\epsilon, b(1 - \psi_\epsilon))$$



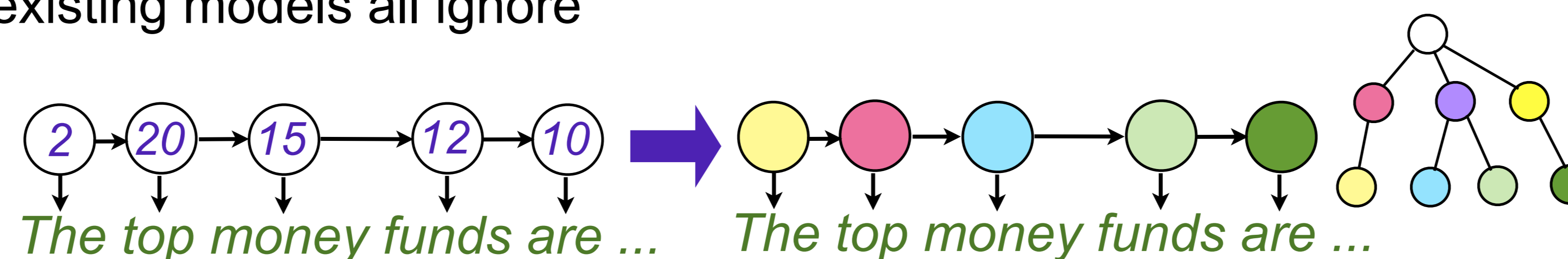
Difference between this ψ -breaks and the HDP

- Let $\varphi_{\epsilon i}^{(k)} = \psi_{\epsilon i}^{(k)} \prod_{j=1}^{i-1} (1 - \psi_{\epsilon j}^{(k)})$ be the local branching prob. to i -th child
- If we model this branching process by HDP, $(\varphi_{\epsilon,1}^{(k)}, \varphi_{\epsilon,2}^{(k)}, \dots) \sim \text{Dir}(b\varphi_{\epsilon,1}, b\varphi_{\epsilon,2}, \dots)$ holds, but in our model, it doesn't
- In HDP, the ψ -break is: $\psi_{\epsilon, i}^{(k)} \sim \text{Beta}(b\varphi_{\epsilon, i}, b(1 - \sum_{j=1}^i \varphi_{\epsilon, j}))$
- Recently proposed nestedCRF [2] is based on HDP; ours is not

HMM on a Tree

Motivation: We want to induce the latent hierarchy of states

- In natural language processing, HMM or other probabilistic grammar models are used to induce word categories for dimensionality reduction
- The word categories should comprise a hierarchical structure, which existing models all ignore

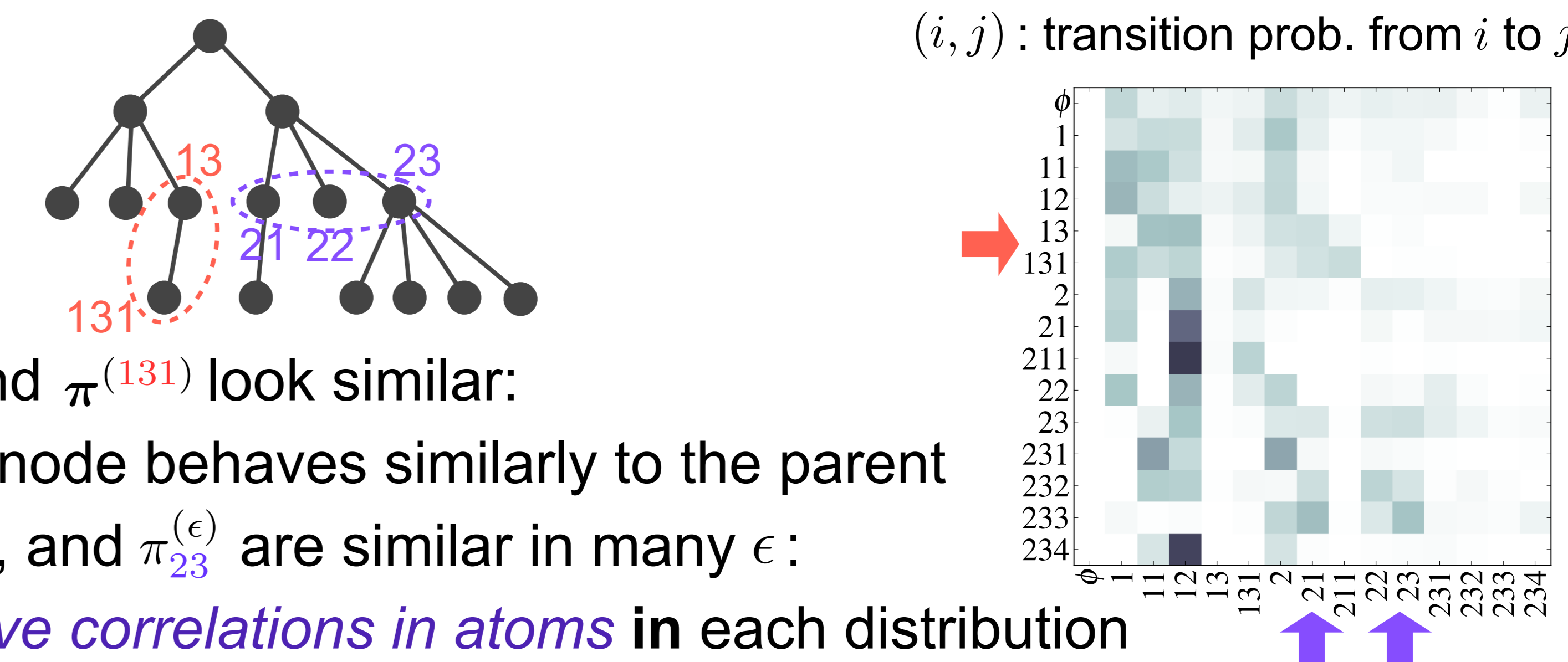


- **Scientific question:** How words are categorized in a tree?
- **Engineering:** the depth of the predicted state corresponds to the *confidence of that state*
- **Assumption:** Two related categories are close to each other
- Each node (category) has *a transition distribution* to other nodes

Generative process

1. Sample $\pi \sim \text{TSSB}(\alpha, \gamma)$ to define the global tree structure
2. On each node $\epsilon \cdot i$, sample $\pi^{(\epsilon \cdot i)} \sim \text{HTSSB}(a, b, \pi^{(\epsilon)})$

A draw from the HMM prior



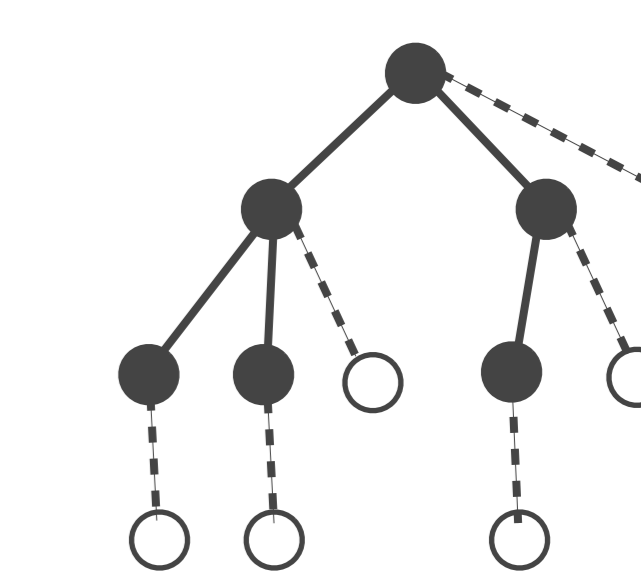
- $\pi^{(13)}$ and $\pi^{(131)}$ look similar:
- Child node behaves similarly to the parent
- $\pi_{21}^{(\epsilon)}$, $\pi_{22}^{(\epsilon)}$, and $\pi_{23}^{(\epsilon)}$ are similar in many ϵ :
- *Positive correlations in atoms in each distribution*

Inference

Gibbs sampler similar to the HDP-HMM:

$$p(z_t | \mathbf{z}^{-t}, \mathbf{w}) \propto p(z_t | z_{t-1}) p(w | z_t) p(z_{t+1} | z_t)$$

To grow the tree, we place dummy nodes (like the dummy state of HDP-HMM)

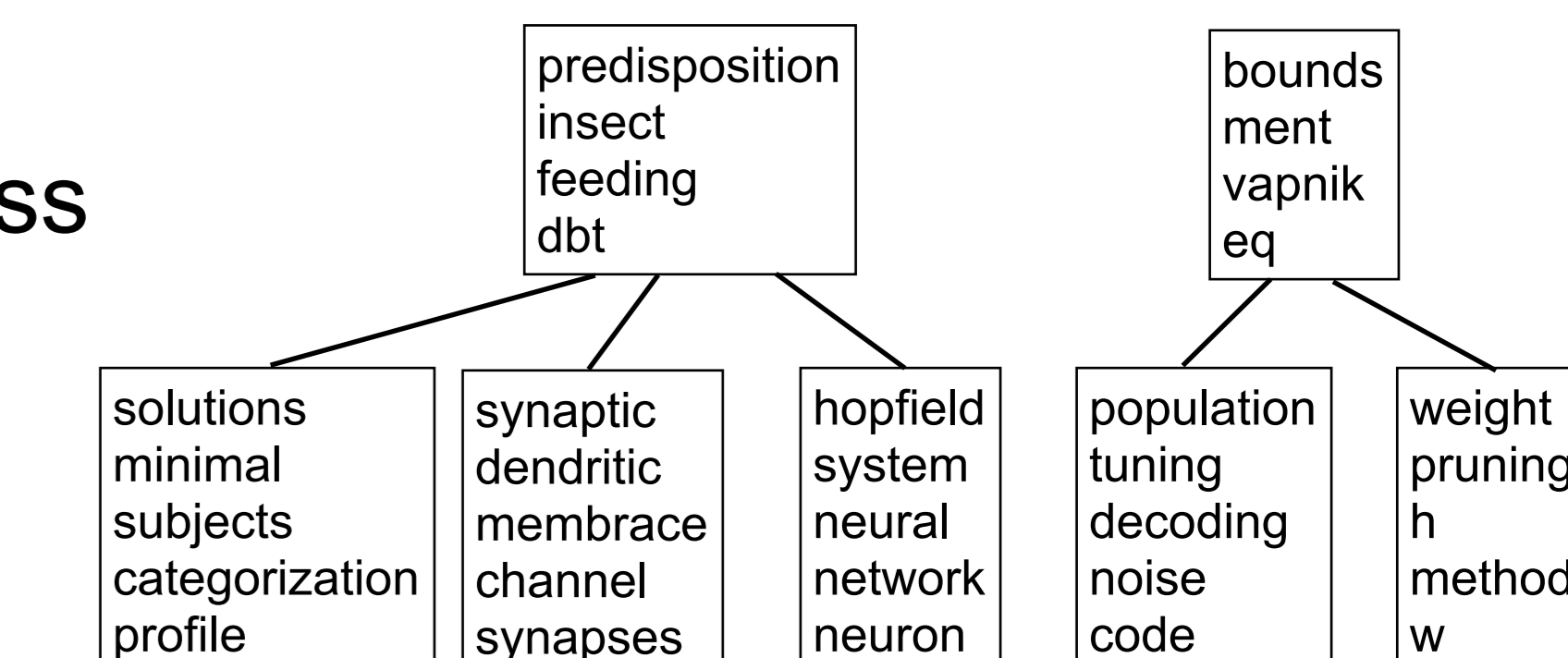


- Currently, this sampler for HMM doesn't work well
- Similar word categories often appear in very different positions
- because the effect of an ancestor diminishes in deeper nodes:

$$\pi^{(\epsilon \cdot i)} \sim \text{HTSSB}(a, b, \pi^{(\epsilon)}); \pi^{(\epsilon \cdot i \cdot j)} \sim \text{HTSSB}(a, b, \pi^{(\epsilon \cdot i)}); \dots$$

Tree-structured Topic Modeling: $\pi^{(d)} \sim \text{HTSSB}(a, b, \pi)$

- It is easier than the HMM, so we can check the correctness of the model and sampler
- From the NIPS corpus, we got reasonable subtrees \Rightarrow



Discussion

- For HMM to work, we need to solve several problems:
 - A blocked sampler, which enable larger moves, might be required
 - Theoretical analysis of the behavior *with deeper hierarchy*
- Interesting applications of Tree-HMM in other domains?

Reference

- [1] R.P. Adams, Z. Ghahramani, and Michael I Jordan. Tree-structured stick breaking for hierarchical data. In *NIPS* 2010
- [2] Amr Ahmed, Liangjie Hong, and Alexander Smola. Nested chinese restaurant franchise process: Applications to user tracking and document modeling. In *Proc. of ICML* 2013