University of Amsterdam

Masters Thesis

# Automated Generation of Textual Categorization Trees Based on Hierarchical Topic Modeling

*Author:*
Wenlin Chen

*Supervisor:*
Yongxin Tang
Sandjai Bhulai

*A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Computational Science*

*in the*

Computational Science Lab
Informatics Institute

August 2025

# Contents

**Declaration of Authorship** i

**Abstract** iii

**Acknowledgements** iv

**Contents** v

**List of Figures** vii

**List of Tables** ix

**List of Algorithms** x

**Abbreviations** xi

**1 Introduction** 1

**2 Literature Review** 3

**3 Methods** 6

3.1 Bayesian Foundation . . . 6

3.1.1 Prior, Observed data, Posterior . . . 6

3.1.2 Conjugacy and the Choice of Dirichlet Priors . . . 7

3.2 From Dirichlet Distribution to Dirichlet Process . . . 8

3.2.1 From Beta to Dirichlet: A Natural Generalization . . . 8

3.2.2 From Dirichlet to Dirichlet Process: Finite to Infinite . . . 9

3.2.3 Stick-breaking Construction of the DP . . . 10

3.2.4 Posterior Conjugacy and the Chinese Restaurant Process . . . 10

3.3 The Nested Chinese Restaurant Process . . . 12

3.4 Gibbs Sampling . . . 13

3.4.1 Path Resampling for Documents . . . 13

3.4.1.1 Likelihood Term Derivation . . . 13

3.4.1.2 Prior Term (nCRP) . . . 14

3.4.2 Sampling Level Allocations (Words to Levels) . . . 15

3.4.2.1 Topic Allocation Probability $z_{d,n} = j$ . . . . . 15
3.4.2.2 Word Generating Probability $w_{d,n}$ . . . . . 16
3.5 Convergence of Gibbs Sampling in hLDA . . . . . 17
3.6 Experimental Design for Hyperparameter Optimization . . . . . 19
3.6.1 Basic Convergence Monitor Indicator . . . . . 20
3.6.2 Cross Multi-chains Convergence Monitor . . . . . 21
3.6.3 Three-Stage Grid Search Framework with Selection Criteria . . . . 22
3.6.3.1 Stage 1: explore tree depth and $\gamma$, fix $\eta$ and $\alpha$ . . . . . 24
3.6.4 Feasibility and Evaluation . . . . . 25
3.6.4.1 Rényi Entropy . . . . . 25
3.6.4.2 Jensen-Shannon Divergence . . . . . 26
3.6.4.3 Perplexity . . . . . 27
3.6.4.4 Coherence . . . . . 27
3.6.4.5 Balance degree of tree structure . . . . . 29
3.7 Comparative Topic Models . . . . . 30
3.7.0.1 Latent Dirichlet Allocation (LDA) . . . . . 30
3.7.0.2 Correlated Topic Model (CTM) . . . . . 30
3.7.0.3 Hierarchical Dirichlet Process (HDP) . . . . . 30

**4 Experiments and Results Analysis 31**
4.1 Data Preparation . . . . . 31
4.1.1 Data Cleaning . . . . . 31
4.2 Experimental Setup and Evaluation Protocol . . . . . 32
4.3 Stage 1: explore tree depth and $\gamma$, fix $\eta = 0.1$, $\alpha = 0.1$ . . . . . 34
4.3.1 Stage 2: Tuning Topic-Word Prior $\eta$ . . . . . 39
4.3.2 Stage 3: Optimizing Topic Diversity with $\alpha$ . . . . . 43
4.4 Model Comparation . . . . . 44
4.5 Experiment Result Analysis . . . . . 46

**5 Discussion 49**

**6 Conclusion and future work 52**

**Bibliography 54**

# List of Figures


3.1 This figure from "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies" by Blei et al.[1], showing a three-level nested Chinese restaurant process (nCRP). Blue squares represent topic nodes (restaurants), yellow diamonds are tables (subtopics), and five customers follow four unique paths, illustrating hierarchical topic clustering. . . . . 12

3.2 Convergence analysis showing the Jaccard Mean (blue line) and Average Stable Documents Count (red dashed line) increase versus iteration, with Jaccard Mean stabilizing around 0.93 and Average Stable Docs Count rising to approximately 580 by 400 iterations . . . . . 21

3.3 Diagram of a three-step grid search approach inspired by topic modeling practices, where Stage 1 initializes with depth and $\gamma$ , Stage 2 processes with $\eta$, and Stage 3 optimizes $\alpha$ for an Optimal Combination. A series evaluation metrics are taken in each stage. . . . . 23

4.1 Analysis of Renyi Entropy and JSD across different model depths and gamma values. The plots on the left show the Weighted Average R . . . . 36

4.2 The plots compare weighted average entropy (left) and distinctiveness JSD (right) across layers for models with depths of 3, 4, and 5 under different gamma ($\gamma$) settings. When depth=3, the consistent decrease in entropy and increase in JSD with layer depth demonstrate that the model effectively learns a hierarchy of increasingly specific and distinct topics . . 38

4.3 Violin plots of word counts per node across layers under different Gamma parameters. With the model depth fixed at 3, the figure shows the distribution of total words per node for each layer (Layer 0, 1, 2) as Gamma varies from 0.001 to 0.1. The plots clearly show that the root node (Layer 0) contains the vast majority of words, while the word count distribution (median, quartiles) in deeper layers (Layer 1, 2) adjusts accordingly with the changing Gamma value . . . . . 40

4.4 Radar chart comparing the performance of different Eta parameters (from 0.005 to 0.2) on six key metrics. The chart evaluates Perplexity, NPMI, C_V, JS Distance, Renyi Entropy, and Gini coefficient. All metrics are normalized and unified to a 'higher is better' standard. The results show that eta=0.05 (the red line) exhibits strong overall performance . . . . . 41

4.5 Radar chart comparing the performance of different alpha parameter values across eight evaluation metrics. All metrics are normalized so that a higher value (closer to the outer edge) indicates better performance. The analysis shows that an alpha value of 0.1 (red line) provides the most robust and balanced performance, achieving high scores on topic coherence (NPMI, C_V), topic distinctiveness (Weighted JSD, Simple Layer JS), and perplexity. . . . . 43

4.6 Visualization of a sub-tree from the hLDA-generated hierarchy, optimized with parameters (depth = 3, $\gamma = 0.05$, $\eta = 0.05$, $\alpha = 0.1$). Green nodes indicate stable nodes consistent across the past 5 rounds, while orange nodes represent unstable ones. Derived from 970 *Computational Mechanics* abstracts (2016–2022), it reflects domains from "numerical modeling" to specialized topics like "finite element formulation." The pink box highlights the stable-to-total document ratio in persisted nodes. . . . 47

# List of Tables

4.1 Example entry from the cleaned corpus. . . . . 32
4.2 A series of metrics of structural and distribution prior fields are adopted to evaluate hLDA model performance. . . . . 34

# Chapter 1

# Introduction

The rapid expansion of digital scientific literature has heightened the demand for robust document organization and information retrieval methods. Research abstracts, though brief, are semantically rich, encapsulating complex concepts in concise form, making them both valuable and challenging for knowledge structuring. In fields like medicine and computer science, established taxonomies such as MeSH and ACM CCS streamline search and classification. However, in physics and related applied sciences, the absence of a standardized hierarchical taxonomy complicates efficient literature retrieval, hindering both researchers and automated systems in navigating and indexing relevant work.

Topic modeling provides a data-driven approach to uncover latent structures within corpora. Flat models like Latent Dirichlet Allocation (LDA) identify hidden themes but fail to capture the hierarchical progression from broad disciplines to specific subfields inherent in scientific knowledge. Hierarchical Latent Dirichlet Allocation (hLDA), built on the Nested Chinese Restaurant Process (nCRP), addresses this by constructing hierarchical topic trees. Higher-level topics reflect broad disciplines, while deeper nodes represent specialized subtopics, creating an interpretable categorization framework well-suited for scientific information retrieval.

Despite its potential, hLDA presents challenges. Its nonparametric nature demands meticulous hyperparameter tuning, and its stochastic inference process complicates convergence assessment. Traditional metrics like log-likelihood often fail to align with semantic stability, while coherence measures offer limited insight. This raises two key questions: How can hyperparameters for hLDA be systematically optimized? And how can convergence be evaluated to ensure both statistical and semantic stability?

This thesis proposes two methodological advancements to address these issues. First, a three-stage hyperparameter search strategy is introduced, sequentially optimizing tree

depth and branching factor ($\gamma$), topic–word prior ($\eta$), and document–path prior ($\alpha$). This approach separates structural and distributional tuning, enhancing interpretability and computational efficiency. Second, a novel convergence monitoring framework is developed, assessing stability through document assignment consistency using the R-hat statistic across parallel chains, complemented by Jaccard similarity of document-path distributions. These innovations establish a robust and interpretable foundation for applying hLDA to scientific texts.

Applied to a corpus of approximately one thousand abstracts from Computational Mechanics (2016–2022), this framework demonstrates hLDA's potential as a surrogate classification system in domains lacking standardized taxonomies. Beyond methodological contributions, this work paves the way for improved information retrieval in physics-related fields, where data-driven hierarchies can substitute for missing ontologies, enabling researchers to navigate the literature more effectively.

# Chapter 2

# Literature Review

In an era where digital texts—from scientific abstracts to industry reports—proliferate at an unprecedented scale, the ability to automatically uncover meaningful structures within vast, unstructured corpora has become a cornerstone of knowledge discovery. Yet, real-world text data, particularly concise and semantically dense documents like research abstracts, poses significant challenges. These texts often span disciplinary boundaries, demand models that can infer latent topics without labeled data, and require computational efficiency to handle large-scale datasets. Traditional flat classification methods, which treat categories as independent entities, struggle to capture the nested relationships inherent in such data, where broad fields like biology branch into specialized subtopics like genomics [2]. Meanwhile, supervised methods falter when faced with sparse or uneven labels, a common issue in real-world corpora. This has driven the field toward unsupervised hierarchical learning, with models like hierarchical Latent Dirichlet Allocation (hLDA) emerging as a powerful solution to infer both the number and organization of topics in a flexible, nonparametric framework [3].

The journey of text classification reflects a broader evolution in handling complex semantic structures. Early flat models, such as Latent Dirichlet Allocation (LDA), treated documents as mixtures of topics and topics as mixtures of words [4]. While LDA sparked enthusiasm for automatic topic discovery, its inability to capture hierarchical relationships limited its effectiveness for layered knowledge structures. To address this, supervised and semi-supervised variants emerged. Hierarchically Supervised LDA (HSLDA) leverages labeled data to tie documents to specific paths, improving predictions in domains like medical records [5]. Similarly, Semi-Supervised hLDA (SSHLDA) uses partial labels to guide topic discovery [6], while Hetero-Labeled LDA (hLLDA) integrates diverse label types, such as document tags and word hints, to handle incomplete expert

annotations [7]. These models excel in scenarios with prior knowledge but struggle in fully unsupervised settings where labels are absent or impractical to obtain.

The limitations of label-dependent approaches become stark as data volumes grow and new domains emerge. Human labeling is labor-intensive, time-consuming, and often plagued by semantic ambiguity, making it infeasible for large-scale or dynamic datasets. Moreover, knowledge hierarchies are increasingly complex, exhibiting inheritance relationships that demand models capable of more than simple categorization. Bayesian nonparametric methods, such as the Nested Chinese Restaurant Process (nCRP) integrated into hLDA, offer a breakthrough by allowing documents to select topic paths from root to leaf without predefined topic counts or layer numbers [4]. In hLDA, parameters like depth (controlling hierarchy levels) and $\gamma$ (influencing branching) shape a tree that grows organically, reflecting the data's intrinsic structure. This flexibility makes hLDA ideal for applications like scientific literature analysis, where parent–child relationships mirror real-world taxonomies.

Despite its strengths, hLDA's nonparametric and unsupervised nature introduces challenges, particularly in hyperparameter tuning and model evaluation. Without labels, assessing the meaningfulness of inferred structures is a persistent hurdle. Traditional metrics like perplexity, defined as the exponential of the negative normalized log-likelihood, measure predictive power but often correlate poorly with human-judged interpretability [4, 8]. Topic coherence metrics, such as Pointwise Mutual Information (PMI) and Normalized PMI (NPMI), improve alignment with human judgment by quantifying the semantic relatedness of top words within a topic [9]. However, these metrics rely on external corpora and do not directly evaluate hierarchical structures. To address this, structure-aware metrics have emerged. Hierarchical Loss (H-loss) penalizes errors based on their depth in the hierarchy, while Tree-based Precision and Recall assess path accuracy via the Lowest Common Ancestor (LCA) [10]. Tree Edit Distance quantifies structural differences by measuring operations needed to align hierarchies, making it ideal for comparing models under different hyperparameter settings [2].

Recent advances in information-theoretic approaches offer further promise. Metrics like Rényi entropy and Jensen–Shannon distance (JSD) capture topic stability and informativeness, reducing the impact of rare words exaggerated by Dirichlet smoothing [11]. These tools help identify the "sweet spot" for hierarchical structures, particularly in unsupervised settings where labeled ground truth is unavailable. Additionally, convergence diagnostics like the Gelman–Rubin statistic (R-hat) ensure model stability by comparing variance across multiple chains, preventing premature stopping [12]. Other methods, such as path consistency and entropy-based stability indicators, assess when document

and word allocations stabilize . These advances underscore the need for a robust evaluation framework that balances statistical rigor with semantic interpretability.

Building on these insights, my research proposes a novel framework to enhance hLDA's robustness, efficiency, and interpretability for complex, real-world texts. I adopt a three-phase hyperparameter tuning strategy to address the computational overhead and instability of stochastic path assignments. First, I refine structural parameters like *depth* and $\gamma$ to establish a balanced tree. Next, I optimize $\eta$, which governs word probability concentration to enhance semantic clustering. Finally, I fine-tune $\alpha$ to ensure even word distribution across topics along a document's path. This "structure-first, semantics-later" approach minimizes inefficient exploration and improves interpretability. To evaluate the resulting hierarchies, I integrate Rényi entropy, JSD, and R-hat, providing real-time insights during training for timely adjustments and early stopping. This framework not only assesses the quality and coherence of the final topic hierarchy but also streamlines the model-building process, offering a scalable solution for automated knowledge organization in domains like scientific abstracts and product ontologies. By addressing these challenges, my work aims to advance hierarchical topic modeling, contributing new theoretical and practical tools to navigate the complexities of unstructured, label-free data.

# Chapter 3

# Methods

## 3.1 Bayesian Foundation

### 3.1.1 Prior, Observed data, Posterior

Our model is grounded in the Bayesian framework, which depicts uncertainty through probability distributions. Unlike frequentist statistics, which interprets probability as the long-run frequency of events and treats model parameters as fixed but intuitive values, Bayesian statistics treats parameters themselves as random variables. A key distinguishing feature is the introduction of a prior distribution, which encodes our initial beliefs or assumptions about a parameter before observing any data. These priors can be empirically informed, based on domain knowledge, or simply uninformative, such as a uniform distribution, when no prior information is available.

To illustrate the philosophical difference, consider the case of flipping a coin. A frequentist approach assumes—often implicitly based on empirical tradition—that the coin has exactly two sides, and estimates the probability of heads or tails based on observed frequencies. A Bayesian, however, might begin with a prior hypothesis that accounts for a possible third outcome, such as the coin landing on its edge. This belief is then updated in light of data: if only heads and tails are ever observed, the posterior will reflect diminishing probability for the third outcome—but crucially, the possibility was encoded from the start. The classical Bayes' theorem is given in Eq: 3.1.

$$\underbrace{P(\theta|D)}_{\text{Posterior}} \propto \underbrace{P(D|\theta)}_{\text{Likelihood}} \cdot \underbrace{P(\theta)}_{\text{Prior}} \tag{3.1}$$

In the context of text analysis, Bayes' theorem allows us to infer the latent structure—such as topics—underlying a collection of documents by combining prior assumptions about topic distributions with the likelihood derived from observed words.

### 3.1.2 Conjugacy and the Choice of Dirichlet Priors

In Bayesian modeling, we often face the challenge of performing inference when the posterior distribution has no closed-form expression. A pragmatic solution is to select a conjugate prior for a given likelihood, meaning that the prior and posterior belong to the same distributional family. This choice is not an inherent property of any single distribution, but rather a feature of the relationship between the prior and the likelihood functions.

In our research scenario, both the topic–word distributions and document–topic distributions are naturally modeled as multinomial over a vocabulary of words or a set of latent topics. Specifically: Each topic is characterized by a multinomial distribution over the vocabulary, representing the probability of each word appearing under that topic; Each abstract is modeled by a multinomial distribution over the set of inferred topics, indicating the proportion of each topic in the document.

Given this multinomial structure, we select the Dirichlet distribution as the conjugate prior for both types of distributions. The mathematical proof of conjugacy relationship between multinomial and Dirichlet is shown as below.

Consider a multinomial likelihood over $K$ categories with observed count vector $\mathbf{n} = (n_1, \ldots, n_K)$ and total trials $N = \sum_k n_k$, the likelihood probability is shown as below; where $\boldsymbol{\theta}$ is our target topic-words distribution and $\theta_k$ represents the probability to generating word $k$ from topic $\theta$.

$$P(\mathbf{n} \mid \boldsymbol{\theta}) = \frac{N!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \theta_k^{n_k} \propto \prod_{k=1}^{K} \theta_k^{n_k}.$$

Let the prior over $\boldsymbol{\theta}$ be the Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, where $\boldsymbol{\alpha}$ is the concentration parameter to control the contribution of our prior brief, each $\alpha$ interpretable as prior pseudo-counts for each category. A Larger $\alpha_k$ encourages more weight on category $k$, while a smaller $\alpha_k$ encourages balanced mixtures from observation.

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K), \quad P(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}.$$

The unnormalized posterior is then:

$$P(\boldsymbol{\theta} \mid \mathbf{n}, \boldsymbol{\alpha}) \propto \left(\prod_{k=1}^{K} \theta_k^{n_k}\right)\left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right) = \prod_{k=1}^{K} \theta_k^{n_k + \alpha_k - 1}.$$

This is exactly the kernel of a Dirichlet distribution with updated parameters:

$$\boldsymbol{\theta} \mid \mathbf{n} \sim \mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K).$$

## 3.2 From Dirichlet Distribution to Dirichlet Process

In the preceding section, we introduced the Dirichlet distribution as the conjugate prior to multinomial likelihoods, emphasizing its role in Bayesian inference and Gibbs sampling. In this section, we extend our exposition by detailing how the Dirichlet distribution naturally generalizes into the Dirichlet Process (DP), a foundational construct enabling Bayesian nonparametric modeling, and essential to our hierarchical topic modeling framework.

### 3.2.1 From Beta to Dirichlet: A Natural Generalization

We begin with the Beta distribution, the simplest Dirichlet-family distribution, defined over a single probability variable $x \in [0, 1]$ with two positive parameters $\alpha > 0$ and $\beta > 0$ and the density function is shown in Eq: 3.2.

$$\mathrm{Beta}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha - 1}\,(1 - x)^{\beta - 1}, \quad 0 \le x \le 1. \tag{3.2}$$

Serving as the conjugate prior to Bernoulli and binomial likelihoods, the Beta distribution's parameters $\alpha, \beta$ can be interpreted as prior pseudo-counts representing "successes" and "failures" respectively. After observing $n$ trials with $k$ successes and $n - k$ failures, the posterior of Beta distribution is:

$$\mathrm{Beta}(x \mid \alpha + k,\ \beta + n - k).$$

While Beta Distribution has two categories - success and failure, Dirichlet distribution has K outcomes. Generalizing from two outcomes (success and failure) to $K$ outcomes leads directly to the Dirichlet distribution. A random vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, defined

on the $(K-1)$-simplex: $\theta_k \geq 0$, $\sum_k \theta_k = 1$ is Dirichlet-distributed if :

$$\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where parameters $\alpha_k > 0$ again represent pseudo-counts reflecting prior knowledge about the distribution across categories. Given multinomial observations $(n_1, \ldots, n_K)$, the posterior updates to a new Dirichlet distribution:

$$\mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1 + n_1, \ldots, \alpha_K + n_K).$$

### 3.2.2 From Dirichlet to Dirichlet Process: Finite to Infinite

Despite its advantages, the Dirichlet distribution has a critical limitation: it requires the number of categories $K$ to be fixed beforehand. However, many real-world problems, such as clustering or topic modeling in evolving text corpora, demand flexibility in determining the number of categories. To address this, we generalize the Dirichlet distribution to the Dirichlet Process (DP), which defines a distribution over distributions, allowing for an unbounded number of components.

Formally introduced by Ferguson (1973), the DP is a distribution over probability measures. Given a measurable space $\Theta$, a base distribution $H$ on $\Theta$, and a concentration parameter $\alpha > 0$, a random probability measure $G$ follows a Dirichlet Process, denoted by $G \sim \mathrm{DP}(\alpha, H)$, if for any finite measurable partition $\{A_1, \ldots, A_r\}$ of $\Theta$:

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_r)).$$

Since for the category-vector $\boldsymbol{\theta} : (\theta_1, \ldots, \theta_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$, the average probability for any $\theta_i$ is :

$$\mathbb{E}[X_i] \;=\; \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}.$$

Intuitively, the base distribution $H$ defines the prior expectation of $G$ since it holds that,

$$\mathbb{E}\big[G(A_i)\big] \;=\; \frac{\alpha\, H(A_i)}{\sum_{k=1}^{r} \alpha\, H(A_k)} \;=\; \frac{\alpha\, H(A_i)}{\alpha \sum_{k=1}^{r} H(A_k)} \;=\; H(A_i).$$

The concentration parameter $\alpha$ controls how closely samples from $G$ resemble $H$. When $\alpha$ is large, realizations of $G$ closely approximate the base distribution $H$. Conversely, when $\alpha$ is small, samples from $G$ tend to concentrate on fewer atoms, facilitating pronounced clustering. In the extreme, as $\alpha \to 0$, a draw from $\mathrm{DP}(\alpha, H)$ collapses all mass to a single atom, resulting in maximum clustering. As $\alpha \to \infty$, $G$ approaches $H$.

A fundamental property of the DP is that its realizations are discrete almost surely, regardless of whether $H$ is continuous. Thus, any $G \sim \mathrm{DP}(\alpha, H)$ can be represented as a discrete sum of atoms:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \quad \text{with} \sum_{k=1}^{\infty} \pi_k = 1,$$

where $\delta_{\theta_k}$ is a point mass at $\theta_k \in \Theta$. This discrete nature facilitates clustering: independently and identically distributed samples from $G$ have positive probability of repeating, naturally forming clusters around shared atoms. The number of clusters is thus allowed to grow adaptively, guided by the concentration parameter $\alpha$.

### 3.2.3 Stick-breaking Construction of the DP

A constructive representation of a DP measure is provided by Sethuraman's stick-breaking process (Sethuraman, 1994). It explicitly generates random weights $\pi_k$ which defines a discrete distribution over atoms $\theta_k$, ensuring $\sum_k \pi_k = 1$.

$$V_k \sim \mathrm{Beta}(1, \alpha), \quad \pi_k = V_k \prod_{j=1}^{k-1} (1 - V_j), \quad \theta_k \sim H.$$

Intuitively, this process corresponds to sequentially breaking a unit-length stick into segments whose lengths are determined by independent Beta draws. This construction not only clarifies the discrete structure of $G$ but also provides a practical basis for variational inference and truncated approximations, where only a finite subset of atoms and weights are considered explicitly.

### 3.2.4 Posterior Conjugacy and the Chinese Restaurant Process

A fundamental advantage of the Dirichlet Process is its conjugacy, meaning the posterior distribution after observing data remains a Dirichlet Process. Initially, we specify a prior distribution with base measure $H$ and concentration parameter $\alpha$. As observations are made, this prior is dynamically updated, with the empirical data informing the posterior distribution.

Formally, suppose we observe samples $\theta_1, \ldots, \theta_n \sim G$. For any finite measurable partition $\{A_1, \ldots, A_r\}$ of the parameter space $\Theta$, let $n_j$ denote the number of observations falling within region $A_j$:

$$n_j = \sum_{i=1}^{n} \mathbb{I}(\theta_i \in A_j).$$

By definition of the Dirichlet Process, the prior distribution of masses for this partition is:

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)).$$

Given the multinomial likelihood of the observed counts $n_j$, and leveraging the conjugacy between Dirichlet and multinomial distributions, the posterior distribution becomes:

$$(G(A_1), \ldots, G(A_r)) \mid \theta_{1:n} \sim \mathrm{Dir}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_r) + n_r).$$

To generalize this finite-dimensional posterior to the infinite-dimensional Dirichlet Process posterior, we recognize that this updated distribution over partitions must correspond to a Dirichlet Process with updated parameters. Indeed, the posterior distribution of the Dirichlet Process is expressed explicitly as:

$$G \mid \theta_{1:n} \sim \mathrm{DP}\left(\alpha + n,\ \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right).$$

Here, the posterior concentration parameter $\alpha + n$ reflects an increase in precision as more data accumulate, and the updated base measure is a weighted combination of the original prior measure $H$ and the empirical distribution formed by the observed data points.

By integrating out the random measure $G$, we derive the predictive distribution for a new observation $\theta_{n+1}$:

$$\theta_{n+1} \mid \theta_{1:n} \sim \frac{\alpha}{\alpha + n} H + \sum_{k=1}^{K} \frac{n_k}{\alpha + n} \delta_{\phi_k},$$

where $\phi_k$ represent the distinct atoms observed so far, with corresponding frequencies $n_k$. This conditional predictive distribution is precisely the Chinese Restaurant Process (CRP). In the CRP analogy, each new data point (or "customer") joins an existing cluster ("table") with probability proportional to its current cluster size, or initiates a new cluster with probability proportional to the concentration parameter $\alpha$:

$$P(\theta_{n+1} = \phi_k \mid \theta_{1:n}) = \frac{n_k}{\alpha + n}, \quad P(\theta_{n+1} \text{ is new}) = \frac{\alpha}{\alpha + n}.$$

The CRP is thus the marginal distribution over cluster assignments induced by integrating out $G \sim \mathrm{DP}(\alpha, H)$, and provides an intuitive generative mechanism to dynamically adjust cluster assignments from prior initialization to posterior updates as data continue to accumulate.

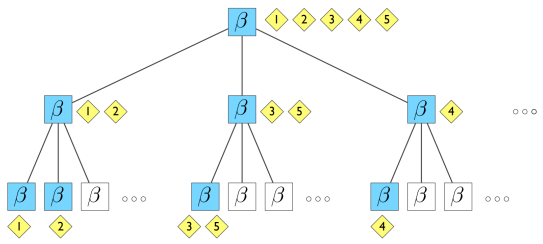


FIGURE 3.1: This figure from "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies" by Blei et al.[1], showing a three-level nested Chinese restaurant process (nCRP). Blue squares represent topic nodes (restaurants), yellow diamonds are tables (subtopics), and five customers follow four unique paths, illustrating hierarchical topic clustering.

## 3.3 The Nested Chinese Restaurant Process

To extend the flexibility of the Dirichlet Process (DP) from flat clustering to hierarchical clustering, we employ the Nested Chinese Restaurant Process (nCRP) introduced by Blei et al. The nCRP places a nonparametric prior on infinitely deep, infinitely branching trees, so that each document $d$ selects a path $(\text{root}, c_{d,1}, c_{d,2}, \ldots, c_{d,\ell})$ from the root node down to deeper levels' nodes, thereby inducing a hierarchy of topics from coarse to fine.

Each node $p$ in the tree also defines its own probability measure $G_p$ over the next-level topics (child atoms). Concretely, when a new table $k$ is created under node $p$, its associated topic parameter $\phi_{p,k}$ is drawn conditionally from $G_p$. Thus, as a document descends the tree, it successively samples topic parameters, and enables documents to traverse a path from general to specific topics.

$$G_p \sim \mathrm{DP}(\alpha,\, G_{\mathrm{parent}(p)}), \quad \phi_{p,k} \sim G_p.$$

Crucially, the nCRP can be understood as a cascade of independent CRPs, one CRP at each node in the tree. Below is a single, level-indexed CRP update rule using $c_{d,\ell} = k$ to denote that document $d$ chooses child $k$ at level $\ell$, and $\alpha$ as the concentration parameter:

$$P\big(c_{d,\ell} = k \mid c_{d,\ell-1} = p\big) = \begin{cases} \dfrac{n_{p,k}}{\alpha + \sum_{k'}^{p} n_{p,k'}} & \text{if child } k \text{ of node } p \text{ is occupied,} \\[2ex] \dfrac{\alpha}{\alpha + \sum_{k'}^{p} n_{p,k'}} & \text{if } k \text{ is a new (unoccupied) child of } j. \end{cases}$$

Here,

$c_{d,0} = \text{root}$.

$n_{p,k}$ is the number of documents previously assigned from node $p$ to its child $k$.

$\sum_{k'}^{p} n_{p,k'}$ is the total number of documents that have passed through node $p$.

$\alpha > 0$ is the global concentration parameter, identical at every level.

## 3.4 Gibbs Sampling

Having established the Nested Chinese Restaurant Process (nCRP) and its role in building hierarchical topic structures, we now turn to inference: specifically, Gibbs sampling for posterior estimation. Gibbs sampling, a Markov Chain Monte Carlo (MCMC) method, iteratively updates the assignment of documents to topic paths and words to levels based on the prior distributions and observed data, eventually converging to a stable, data-informed hierarchical structure.

### 3.4.1 Path Resampling for Documents

In hLDA, each document $d$ is associated with a path $c_d = (c_{d,1}, c_{d,2}, \ldots, c_{d,\ell})$ from the root (level 0) to deeper topic nodes. During inference, the first step of Gibbs sampling involves resampling each document's path conditioned on the current assignments of all other documents. Formally, we sample from:

$$\underbrace{p(c_d \mid \mathbf{w}_d, \mathbf{c}_{-d}, \mathbf{z}, \eta, \alpha)}_{\text{probability over candidate path } c_d} \propto \underbrace{p(\mathbf{w}_d \mid c_d, \mathbf{w}_{-d}, \mathbf{z}, \eta)}_{\text{Likelihood}} \cdot \underbrace{p(c_d \mid \mathbf{c}_{-d}, \alpha)}_{\text{nCRP prior}}$$

Here:

- $\mathbf{w}_d$: Words in document $d$.
- $c_d$: Candidate path for document $d$.
- $\mathbf{c}_{-d}$: Paths for all documents except $d$.
- $\mathbf{z}$: Current topic assignments at each node.
- $\eta$: Hyperparameter for Dirichlet priors on word distributions.
- $\alpha$: nCRP concentration parameter.

#### 3.4.1.1 Likelihood Term Derivation

Under the "bag-of-words" assumption, the words within document $d$ are conditionally independent given their latent topics. Therefore, the likelihood factorizes over levels of the hierarchy:

$$p(\mathbf{w}_d \mid c_d, \mathbf{w}_{-d}, \mathbf{z}, \eta) = \prod_{l=1}^{L} p(\mathbf{w}_d^{(l)} \mid c_d^{(l)}, \mathbf{w}_{-d}^{(l)}, \mathbf{z}^{(l)}, \eta)$$

At each level $l$, the likelihood is obtained by integrating over the latent topic-word distribution $\phi_{c_d^{(l)}}$:

$$p(\mathbf{w}_d^{(l)} \mid c_d^{(l)}, \mathbf{w}_{-d}^{(l)}, \mathbf{z}^{(l)}, \eta) = \int p(\mathbf{w}_d^{(l)} \mid \phi_{c_d^{(l)}}) p(\phi_{c_d^{(l)}} \mid \mathbf{w}_{-d}^{(l)}, \eta)\, d\phi$$

Since $\phi_{c_d^{(l)}}$ follows a Dirichlet posterior given previously observed words: Prior: $\phi_{c_d^{(l)}} \sim \mathrm{Dir}(\eta)$, symmetric Dirichlet with hyperparameter $\eta$.
Posterior given other words:

$$\phi_{c_d^{(l)}} \sim \mathrm{Dir}\left(\eta + n^{(w_1)}_{c_d^{(l)},-d}, \ldots, \eta + n^{(w_W)}_{c_d^{(l)},-d}\right)$$

Integrating analytically yields:

$$p(\mathbf{w}_d^{(l)} \mid c_d^{(l)}, \mathbf{w}_{-d}^{(l)}, \mathbf{z}^{(l)}, \eta) = \frac{\Gamma\left(\sum_w^W (n^{(w)}_{c_d^{(l)},-d} + \eta)\right)}{\prod_w^W \Gamma(n^{(w)}_{c_d^{(l)},-d} + \eta)} \cdot \frac{\prod_w^W \Gamma(n^{(w)}_{c_d^{(l)},-d} + n^{(w)}_{c_d^{(l)},d} + \eta)}{\Gamma\left(\sum_w^W (n^{(w)}_{c_d^{(l)},-d} + n^{(w)}_{c_d^{(l)},d} + \eta)\right)}$$

This step balances prior tree structure (via the nCRP) and document-specific word likelihood under candidate paths.

#### 3.4.1.2 Prior Term (nCRP)

The prior distribution over document paths is governed by the Nested Chinese Restaurant Process (nCRP). During Gibbs sampling, each document $d$ re-samples its path down the hierarchy, conditioning on all other documents' current assignments. Specifically, the probability of selecting child node $k$ at level $l$, given its parent node $p = c_{d,l-1}$, is:

$$p(c_{d,l} = k \mid c_{d,l-1} = p, \mathbf{c}_{-d}, \alpha) = \begin{cases} \dfrac{n_{p,k}}{\alpha + \sum_{k'}^{p} n_{p,k'}}, & \text{if child node } k \text{ already exists,} \\ \dfrac{\alpha}{\alpha + \sum_{k'}^{p} n_{p,k'}}, & \text{if choosing a new (unoccupied) child node.} \end{cases}$$

Here, the notation remains consistent with the previous sections:
$n_{p,k}$: Number of documents previously assigned from parent node $p$ to child node $k$.
$\alpha$: nCRP concentration parameter.

An important property of hLDA is that whenever a document selects a new (previously unoccupied) child node at any intermediate node, the model will generate an entirely new path by sequentially creating new child nodes downwards from this point onward until reaching the pre-defined leaf level $\ell$. Thus, once a new node is chosen at some

intermediate level, all subsequent nodes at deeper levels along that path are also newly created.

Each newly created node initially has no observed data and is thus assigned a topic-word distribution directly from a symmetric Dirichlet prior $\mathrm{Dir}(\eta)$:

$$\phi_{\text{new node}} \sim \mathrm{Dir}(\eta).$$

As the sampling process continues, these newly created nodes will gradually accumulate data and thus update their posterior distributions accordingly. This adaptive, non-parametric nature allows hLDA to dynamically expand its hierarchical topic structure, flexibly capturing emergent semantic granularity within the corpus.

### 3.4.2 Sampling Level Allocations (Words to Levels)

After updating each document's path, the next step assigns each word $w_{d,n}$ in document $d$ to a specific topic-level along its chosen path. Critically, Gibbs sampling operates by first removing the current assignment of the word $w_{d,n}$, and then resampling its topic-level assignment conditioned on the assignments of all other words. In other words, Gibbs sampling inherently leverages the conditional independence structure by temporarily excluding the current word, allowing the model to re-evaluate its topic-level assignment based solely on the remaining observed words and their current allocations.

Formally, this conditional probability for assigning word $w_{d,n}$ to topic-level $l$ given all other assignments is computed as:

$$\underbrace{p(z_{d,n} = j \mid w_{d,n}, c_d, \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}, \alpha, \eta)}_{\text{probability for topic } j}$$
$$\propto \underbrace{p(w_{d,n} \mid z_{d,n} = j, \mathbf{w}_{-(d,n)}, \mathbf{z}_{-(d,n)}, \eta)}_{\text{word generating probability}} \cdot \underbrace{p(z_{d,n} = j \mid z_{-(d,n)}, \Omega)}_{\text{topic allocation probability}}$$

#### 3.4.2.1 Topic Allocation Probability $z_{d,n} = j$

The probability of assigning word $w_{d,n}$ to a certain topic-level $j$ depends on the proportions of topics within the document $d$, captured by a latent multinomial variable $\theta_d$. We thus introduce the latent document-topic proportion variable $\theta_d$, which follows a Dirichlet distribution:

**Prior Distribution:** Before observing any words, we assume a symmetric Dirichlet prior for topic proportions:

$$\theta_d \sim \mathrm{Dir}(\Omega) = \frac{\Gamma(T\Omega)}{\Gamma(\Omega)^T} \prod_{t=1}^{T} X_{d,t}^{\Omega-1}$$

**Posterior Distribution:** After observing other words' assignments (excluding $w_{d,n}$, we update the topic proportion with observed counts:

$$\theta_d \mid z_{d,-w_{d,n}}, \Omega \sim \mathrm{Dir}\big(\Omega + n_{d,1,-w_{d,n}}, \ldots, \Omega + n_{d,T,-w_{d,n}}\big)$$

Therefore, the conditional probability is obtained by integrating out the latent $\theta_d$:

$$\begin{aligned} P(z_{d,n} = j \mid z_{-(d,n)}, \Omega) &= \int P(z_{d,n} = j \mid \theta_d) \cdot P(\theta_d \mid z_{-(d,n)}, \Omega)\, d\theta_d = \mathbb{E}[\theta_{d,j}] \\ &= \frac{n_{d,j,-n} + \Omega}{\sum_{k=1}^{T}(n_{d,k,-n} + \Omega)} \end{aligned}$$

where $n_{d,j,-n}$ is the number of other words in document $d$ (excluding $w_{d,n}$ already assigned to topic-level $j$.

#### 3.4.2.2 Word Generating Probability $w_{d,n}$

Similarly, given the assignment of word $w_{d,n}$ to topic-level $j$, the probability of observing this word is governed by the topic-word distributions $\phi_j$. This distribution also follows a Dirichlet prior:

**Prior Distribution**: Each topic-level's word distribution is initially assigned a symmetric Dirichlet prior with hyperparameter $\eta$:

$$\phi_j \sim \mathrm{Dir}(\eta, \ldots, \eta)$$

**Posterior Distribution**: After observing the current assignment of all other words, we update the posterior distribution:

$$\phi_j \mid \mathbf{w}_{-(d,n)}, \mathbf{z}_{-(d,n)}, \eta \sim \mathrm{Dir}\big(\eta + n_{j,-(d,n)}^{(w_1)}, \ldots, \eta + n_{j,-(d,n)}^{(w_W)}\big)$$

By integrating out $\phi_j$, we obtain the marginal likelihood:

$$P(w_{d,n} \mid z_{d,n} = j, \mathbf{w}_{-(d,n)}, \mathbf{z}_{-(d,n)}, \eta) = \int p(w_{d,n} \mid \phi_j)\, p(\phi_j \mid \mathbf{w}_{-(d,n)}, \mathbf{z}_{-(d,n)}, \eta)\, d\phi_j$$
$$= \frac{n_{j,-(d,n)}^{(w_{d,n})} + \eta}{\sum_{w'=1}^{W} n_{j,-(d,n)}^{(w')} + W\eta}$$

This result expresses the predictive probability of observing word $w_{d,n}$ from topic-level $j$, given the counts of that word previously observed in this topic-level.

**Combined Conditional Probability** Combining both the topic-generating and word-generating terms, the conditional Gibbs sampling update for each word assignment is explicitly given by:

$$p(z_{d,n} = j \mid w_{d,n}, c_d, \mathbf{z}_{-(d,n)}, \mathbf{w}_{-(d,n)}, \Omega, \eta) \propto \underbrace{\frac{n_{d,j,-n} + \Omega}{\sum_{k=1}^{T}(n_{d,k,-n} + \Omega)}}_{\text{topic allocation}} \cdot \underbrace{\frac{n_{j,-(d,n)}^{(w_{d,n})} + \eta}{\sum_{w'=1}^{W} n_{j,-(d,n)}^{(w')} + W\eta}}_{\text{word generation}}$$

This clearly illustrates the essential Gibbs sampling principle: we explicitly remove the current word's assignment, recompute the conditional posterior probabilities based on Dirichlet-multinomial conjugacy, and resample its assignment. By repeatedly applying this step to each word in the corpus, the hierarchical model dynamically refines the latent topic structure, iteratively capturing the underlying semantic patterns in the data.

## 3.5 Convergence of Gibbs Sampling in hLDA

For the hLDA model, the set of latent variables can be explicitly written as follows:

- The document-topic distribution $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$, where $\theta_d$ is the topic distribution for document $d$
- The topic-word distribution $\phi = (\phi_1, \phi_2, \ldots, \phi_K)$, where $\phi_k$ is the word distribution for topic $k$
- The word-level topic assignments $z = (z_{d,1}, z_{d,2}, \ldots, z_{d,N_d})$, where $z_{d,n}$ is the topic (or level) assigned to the $n$-th word in document $d$
- The document path assignments $c = (c_1, c_2, \ldots, c_D)$, where $c_d$ represents the path assigned to document $d$ in the hierarchical tree

Our goal is to sample from the joint posterior distribution iterately till convergence:

$$\pi(X) = p(c, z, \theta, \phi \mid w, \alpha, \eta, \gamma),$$

where the state variable $X = (c, z, \theta, \phi)$ includes path assignments $c$, topic assignments $z$, topic proportions $\theta$, and word distributions $\phi$, and $w$ represents the observed words.

The Gibbs sampling transition kernel is defined as:

$$P(X \to X') = p(x_i' \mid X_{-i}, w, \theta),$$

where each hidden variable $x_i'$ is sampled from its conditional distribution $X_{-i}$ given all other variables.

According to Robert & Casella (2004)[13] and Tierney (1994)[14], a Markov chain converges to its unique stationary distribution if it satisfies the following conditions, independent of the specific model:

**Finite or Countable State Space**: If the state space is finite, convergence holds given irreducibility and aperiodicity. In hLDA, the state space is finite, $c_d$ (document paths) is limited by a fixed *depth*, and $z_{dn}$ (word assignments) is constrained to nodes along the path.

**Irreducibility**: any two states can be reached with non-zero probability. hLDA's nCRP assigns a non-zero probability to each new path due to $\gamma > 0$, and the Dirichlet-multinomial distribution ensures non-zero probabilities for topic assignments with $\eta > 0$, making the chain irreducible.

**Aperiodicity**: the Gibbs sampler has positive probability of resampling the same value for $c_d$ or $z_{d,n}$ as in the previous iteration. Therefore, the chain can remain in the same state with non-zero probability, which eliminates periodicity and ensures aperiodicity.

**Detailed Balance**: for each Gibbs update, the transition kernel is

$$P(X \to X') = p(x_i' \mid X_{-i}, \mathbf{w}, \alpha, \eta, \gamma),$$

where $X_{-i}$ denotes all variables except $x_i$. Then

$$\pi(X)P(X \to X') = p(X \mid \mathbf{w}, \alpha, \eta, \gamma) \cdot p(x_i' \mid X_{-i}, \mathbf{w}, \alpha, \eta, \gamma).$$

By Bayes' rule, this is equal to

$$\pi(X')P(X' \to X),$$

so detailed balance is satisfied. Therefore, $\pi(X)$ is a stationary distribution of the chain.

Therefore, the Gibbs sampling chain in hLDA:

1. Has a finite or countable state space (constrained by *depth* and vocabulary size)
2. Is irreducible (all states are reachable)
3. Is aperiodic (positive self-transition probability)
4. Satisfies detailed balance (via conditional sampling)

By the Markov chain convergence theorem (Doeblin's or Ergodic theorem), the chain converges to its unique stationary distribution:

$$\pi(X) = p(c, z, \theta, \phi \mid w, \alpha, \eta, \gamma),$$

confirming that Gibbs sampling in hLDA theoretically converges to the true posterior distribution.

## 3.6 Experimental Design for Hyperparameter Optimization

The hierarchical Latent Dirichlet Allocation (hLDA) model relies on four hyperparameters that govern its hierarchical structure and inference process:

- *depth*: determines the maximum depth of the categorization tree, balancing broad themes against fine-grained subtopics.
- $\gamma$: the concentration parameter of the nested Chinese Restaurant Process (nCRP) prior, controls the likelihood of creating new branches, with higher values favoring exploration and lower values favoring existing paths.
- $\eta$: the Dirichlet prior word Pseudo for topic-word distribution, shapes the concentration or dispersion of each topic's vocabulary, influencing whether topics are sharp or broad.
- $\alpha$: the Dirichlet prior word Pseudo for document-topic distribution, regulates the diversity of topic assignments within a document's path through the tree. These hyperparameters are optimized through a three-step grid search strategy, detailed below, to achieve a robust and semantically meaningful model configuration.

### 3.6.1 Basic Convergence Monitor Indicator

Exploring convergence in hLDA poses challenges to unsupervised learning, as Gibbs sampling is an MCMC process and introduces randomness in the topic path and word assignments. Unlike supervised learning, which relies on error metrics against labeled data to track convergence; the unsupervised framework lacks such direct indicators, necessitating alternative measures to determine when the topic hierarchy stabilizes.

In the hLDA model, we use the **count of stable documents** that consistently selects the same topic path over a sliding window $W$, such as five consecutive iterations, instead of comparing the single-step stability baseline referenced in [1] which only compares adjacent iterations, making it sensitive to transient flips and blind to periodic oscillations between paths.

We define,

- $D$: the number of documents,
- $t$: the current Gibbs sampling iteration,
- $W$: the sliding window width,
- $p_d^{(t)}$: the topic path of document $d$ at iteration $t$.

For $t \geq W$, document $d$ is *stable* in the window $[t-W+1,\ t]$ if its path remains identical across all $W$ iterations,

$$p_d^{(t-W+1)} = p_d^{(t-W+2)} = \cdots = p_d^{(t)}$$

Then, define the set of stable documents and set size as:

$$\mathcal{S}_W^{(t)} \;=\; \left\{\, d \in \{1, \ldots, D\} \;:\; z_d^{(t-W+1)} = \cdots = z_d^{(t)} \,\right\}$$

$$\left|\mathcal{S}_W^{(t)}\right| \;=\; \sum_{d=1}^{D} \mathbf{1} \cdot \left[\, z_d^{(t-W+1)} = \cdots = z_d^{(t)} \,\right]$$

This approach is based on the intuition that, as iterations progress, more documents converge to stable topic paths with the highest log-likelihood, indicating increasingly consistent path and word-topic assignments.

Additionally, we employ the **Jaccard index** as a supporting measure to assess the similarity between the sets of stable documents in two consecutive iterations. Convergence is confirmed when the stable document count stabilizes and the Jaccard index between

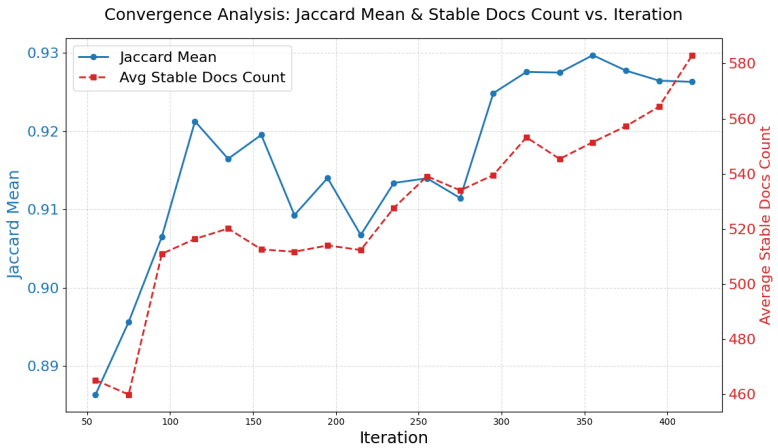


FIGURE 3.2: Convergence analysis showing the Jaccard Mean (blue line) and Average Stable Documents Count (red dashed line) increase versus iteration, with Jaccard Mean stabilizing around 0.93 and Average Stable Docs Count rising to approximately 580 by 400 iterations

these sets exceeds a predefined threshold, for example, 0.95, indicating that a high proportion of documents in the stable set, such as 95%, remain consistently stable across iterations. This threshold can be adjusted based on the specific requirements of the analysis.

The Jaccard similarity between stable sets in consecutive windows [t-1, t] is,

$$J_t = \frac{|S_t \cap S_{t-1}|}{|S_t \cup S_{t-1}|}$$

This captures composition stability but is *not* used alone for convergence due to its insensitivity to the absolute size of $S_t$. It is instead applied as an auxiliary confirmation.

### 3.6.2 Cross Multi-chains Convergence Monitor

After defining the basic indicator for convergence, a challenge emerges when we establish an appropriate threshold for the count of stable documents indicating a stable state, due to limited prior experience in its determination. A viable approach is to compare the stable document counts across multiple chains executed under identical parameter settings. This method assumes that, even in a stochastic process, common characteristics will emerge after sufficient iterations under consistent settings. We employ the **Gelman–Rubin diagnostic** (R-hat), a widely used MCMC convergence statistic that compares within-chain variance and between-chain variance to determine if multiple chains have reached a stationary distribution [**?** ].

R-hat compares the **between-chain variance** $B$ to the **within-chain variance** $W$. If multiple independent chains have mixed to the same stationary distribution, their chain

means should be similar, so $B$ and $W$ should be of comparable magnitude. If chains have not mixed, chain means differ and $B \gg W$. Assuming we run $m$ independent chains, observe the last $n$ draws from each chain and $\theta_{j,t}$ represents the value of the indicator in iteration $t$ from chain $j$.

1. Per-chain mean and variance,

$$\bar{\theta}_j = \frac{1}{n}\sum_{t=1}^{n}\theta_{j,t}, \quad s_j^2 = \frac{1}{n-1}\sum_{t=1}^{n}\left(\theta_{j,t} - \bar{\theta}_j\right)^2, \quad j = 1, \ldots, m.$$

2. Overall mean across chains,

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j$$

3. Within-chain variance **W** (average of per-chain variances),

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2$$

4. Between-chain variance **B** (compares chain means),

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_j - \bar{\theta}\right)^2$$

5. Pooled variance estimate $\hat{\mathrm{Var}}$ is calculated as,

$$\hat{\mathrm{Var}} = \frac{n-1}{n}W + \frac{1}{n}B$$

6. Finally, Gelman–Rubin statistic **R-hat** is,

$$\text{R-hat} = \sqrt{\frac{\hat{\mathrm{Var}}}{W}} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}} = \sqrt{\frac{n-1}{n} + \frac{B}{nW}}$$

When chains converge to the same stationary distribution, their means ($\bar{\theta}_j$) are similar, making the between-chain variance $B$ small and comparable to the within-chain variance $W$, i.e., $B \approx W$ and the R-hat statistic approaches 1. Substituting into the pooled variance:

$$\frac{\hat{Var}}{W} = \frac{n-1}{n} + \frac{B}{nW} \approx \frac{n-1}{n} + \frac{1}{n} = 1, \quad \text{so} \quad \text{R-hat} \approx 1$$

### 3.6.3 Three-Stage Grid Search Framework with Selection Criteria

The hLDA model relies on four hyperparameters, each playing a distinct role in shaping the hierarchical tree structure and influencing model behavior. Prior studies, such as

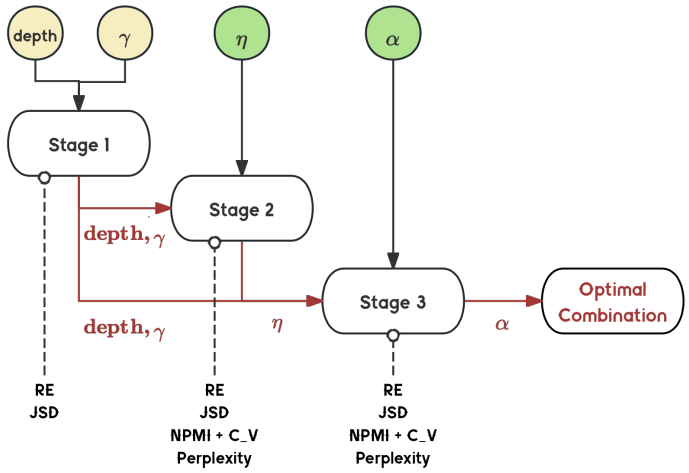


FIGURE 3.3: Diagram of a three-step grid search approach inspired by topic modeling practices, where Stage 1 initializes with depth and $\gamma$ , Stage 2 processes with $\eta$, and Stage 3 optimizes $\alpha$ for an Optimal Combination. A series evaluation metrics are taken in each stage.

Maier et al. (2018) [15] and Egger and Yu (2022)[16], commonly optimize topic model hyperparameters through grid search to identify the optimal configuration. However, each candidate configuration requires sufficient iterations to achieve convergence and yield reliable evaluation metrics, incurring significant computational overhead.We adopt this established practice for the hLDA model, employing a systematic three-step grid search strategy to select optimal hyperparameter settings while minimizing computational load.

This three-step grid search approach is inspired by established practices in topic modeling, where Dirichlet priors are often fixed to focus on structural parameters. For instance, Griffiths and Steyvers (2004)[17] fixed $\alpha$ and $\beta$ in LDA to explore the impact of varying the number of topics, while Li and McCallum (2006)[18] fixed $\alpha$ and $\eta$ in PAM to investigate different DAG structures. Similarly, Mimno et al. (2007) [19] adopted fixed $\alpha$ and $\eta$ values in hPAM to compare hierarchical configurations, and Blei and Lafferty (2007) [20] fixed correlated parameters in CTM to focus on topic count variations.

Moreover, our strategy extends this principle by further separating the tuning of structural *depth* and $\gamma$ and distributional parameters $\eta$ and $\alpha$ into three distinct stages, with an additional step to meticulously optimize $\alpha$, thereby enhancing the granularity of the grid search process.

#### 3.6.3.1 Stage 1: explore tree depth and $\gamma$, fix $\eta$ and $\alpha$

In the first stage, we optimize the structural parameters *depth* and $\gamma$ to establish an effective hierarchical tree structure for the hLDA model. This stage is critical due to the nested Chinese Restaurant Process (nCRP), which defines the initial hierarchy and significantly influences subsequent inference. In the nCRP, the probability of a document selecting an existing branch at a node is proportional to the branch's current document count $n_k$,[1]. While the probability of creating a new branch is proportional to the concentration parameter $\gamma$ [1]. This "rich-get-richer" mechanism introduces a self-reinforcing bias, where early path choices favor existing branches when gamma is small, shaping the tree's topology and constraining the space of possible hierarchies [17].

The parameters *depth* and $\gamma$ serve as inputs to the nCRP, shaping its branching process, while the resulting initial tree acts as an input to Gibbs sampling, impacting the entire model's iterative inference. Further, Gibbs sampling used for hLDA inference exhibits slow mixing across the partition space, struggling to transition between posterior modes [21]. This "stickiness" causes chains to remain close to the initial tree structure, as reassigning a large number of documents to a new branch depends on a series of low-probability updates [22].

Consequently, the initial hierarchy established by *depth* and $\gamma$ has a lasting impact on the posterior distribution, underscoring the need to prioritize their optimization in Stage 1. By carefully selecting these parameters using metrics such as Rényi Entropy and Jensen-Shannon Divergence (JSD) to assess tree structure quality, as discussed in Section - Feasibility and Evaluation, we ensure a robust initial hierarchy that enhances the model's ability to capture meaningful patterns during subsequent reference.

**Stage 1**: Select the optimal tree depth and branching factor ($\gamma$).
**Rationale**: The tree depth and $\gamma$ define the hierarchical structure of the topic model, determining the number of levels and the branching behavior at each node. Optimizing these parameters first establishes a stable structural foundation, ensuring the topic tree aligns with the natural hierarchy of the corpus before refining topic distributions.

**Stage 2**: With the selected *depth* and $\gamma$, tune $\eta$
**Rationale**: As the Dirichlet prior for topic-word distributions, $\eta$ influences topic sparsity and indirectly affects path creation during sampling. Optimizing $\eta$ at this stage refines the vocabulary distribution within topics, enhancing the tree's internal structure without altering its primary framework.

**Stage 3**: Fix *depth*, $\gamma$, and $\eta$, and optimize $\alpha$.
**Rationale**: Since $\alpha$ governs the diversity of topic assignments along a document's

path, tuning it last allows us to balance structural stability with topic diversity. This additional step, compared to prior approaches, ensures a more meticulous optimization by isolating $\alpha$'s effect on document-topic distributions, potentially improving overall model performance.

### 3.6.4 Feasibility and Evaluation

Evaluation metrics are crucial for guiding our three-step grid search strategy to optimize the hLDA model's hyperparameters, ensuring a robust configuration that balances structural integrity, semantic quality, and model fit. In each stage, we apply carefully selected metrics to identify the optimal hyperparameter settings. In Stage 1, with $\eta$ and $\alpha$ fixed, we use Rényi Entropy and Jensen-Shannon Divergence (JSD) to evaluate the hierarchical structure defined by *depth* and $\gamma$. In Stage 2 and Stage 3, we assess the topic-word distribution influenced by $\eta$ and document-topic distribution governed by $\alpha$, we further employing coherence and perplexity to ensure effective semantic classification.

#### 3.6.4.1 Rényi Entropy

Drawing from Koltcov et al. (2020), we analogy topic modeling in the hLDA model as an "information thermodynamics system", where the interplay of order and chaos shapes the clarity of discovered topics [11]. In this framework, the $\phi$ matrix, representing the word-topic distribution, serves as the system's state, capturing the probabilistic landscape of words across topics.

Rényi Entropy , an advanced generalization of Shannon Entropy, is particularly well-suited for evaluating the hLDA model due to its ability to measure the diversity of topic-word distributions, which is critical in the presence of the Dirichlet prior $\eta$'s smoothing effect. For any topic-word distribution $p(w_i)$ Rényi Entropy is defined as:

$$H_q = \frac{1}{1-q} \log \left( \sum_{i=1}^{V} p(w_i)^q \right) \tag{3.3}$$

where $V$ is the vocabulary size, and $p(w_i)$ is the posterior probability of word $w_i$ from its corresponding topic, $q$ is the order coefficient and when

- $q < 1$: Emphasizes rare words, assessing topic diversity in sparse nodes where $\eta$ dominates.
- $q = 1$: Approximates Shannon Entropy $H = -\sum_{i=1}^{V} p(w_i) \log p(w_i)$, balancing all words for overall uncertainty.

- $q > 1$: Focuses on dominant words, ideal for detecting sharp topics in nodes with sufficient documents.

The advantage of Rényi Entropy in our journal abstract datasets is that, topic-word distributions often include high-probability, meaningful terms (e.g., "algorithm") and low-probability noise words (e.g., rare terms). Rényi Entropy with $q > 1$ suppresses noise, focusing on dominant terms to improve topic interpretability, aligning with hLDA's sparse distributions induced by a small $\eta$ [11]. This approach helps assess the quality of topic distributions to achieve clearer and more interpretable topic structures.

#### 3.6.4.2 Jensen-Shannon Divergence

Jensen-Shannon Divergence (JSD) is a critical metric for evaluating topic distinctiveness in the hLDA model by measuring the divergence between topic-word distributions, with higher values indicating greater separation, essential for assessing hierarchical structure in three stages. Since we expect the layer-average JSD to increase as *depth* increases, reflecting enhanced topic differentiation within layers.

For two topics with word distributions $\phi_1 = p(w_i)_{i=1}^V$ and $\phi_2 = q(w_i)_{i=1}^V$, where $V$ is the vocabulary size, JSD is defined as:

$$\mathrm{JSD}(\phi_1, \phi_2) = \frac{1}{2}\mathrm{KL}(\phi_1 \parallel m) + \frac{1}{2}\mathrm{KL}(\phi_2 \parallel m)$$

where $m$ is the compromise distribution of $\phi_1$ and $\phi_2$:

$$m = \frac{1}{2}(\phi_1 + \phi_2)$$

and Kullback-Leibler (KL) divergence calculates the relative entropy of a node's topic-word distribution $\phi_1$ or $\phi_2$ compared to a compromise distribution $m$,

$$\mathrm{KL}(\phi_1 \parallel m) = \sum_{i=1}^{V} p(w_i) \log \frac{p(w_i)}{m(w_i)}.$$

JSD improves upon KL divergence, which is asymmetric and unbounded, by providing a symmetric and bounded measure $0 \leq \mathrm{JSD} \leq \ln 2 \approx 0.6931$ , and it is robust to zero probabilities [**?** ]. In the hLDA model, we compute pairwise JSDs between nodes at a given layer and average them to obtain the layer-average JSD, with higher values indicating greater topic distinctiveness within the layer, reflecting a successful refinement of the hierarchical structure as *depth* increases.

#### 3.6.4.3 Perplexity

To complement structural metrics like Rényi Entropy and Jensen-Shannon Divergence, we introduce perplexity as an additional auxiliary metric to evaluate the classification performance of the hLDA model. Perplexity, a standard metric in language modeling, assesses the predictive accuracy on unseen test data, measuring how well the model assigns topics to new documents. For a test document set $D$, perplexity is defined as:

$$\text{Perplexity}(D) = \exp\left(-\frac{\sum_{d \in D} \log p(w_d)}{\sum_{d \in D} C_d}\right)$$

where $w_d$ represents the words in each document $d$, $p(w_d)$ is the model's estimated probability of these words, and $C_d$ is the count of words in document $d$. Perplexity derives from the total log-likelihood, $\sum_{d \in D} \log p(w_d)$, normalized by the total word count, $\sum_{d \in D} C_d$, and transformed exponentially:

$$\text{Perplexity}(D) = \exp\left(-\frac{\text{Total Log-Likelihood}}{\text{Total Word Count}}\right)$$

Lower perplexity indicates better topic assignment accuracy [4]. In hLDA, perplexity is critical in Stages 2 and 3 of our grid search strategy when optimizing $\eta$ and $\alpha$, because it measures the model's predictive accuracy for topic assignments, ensuring effective classification of documents into hierarchical topics. Perplexity, alongside coherence, enhances classification evaluation, complementing Rényi Entropy and JSD to ensure a robust and interpretable hierarchy.

Perplexity, alongside coherence, enhances classification evaluation, complementing Rényi Entropy and JSD to ensure a robust and interpretable hierarchy.

#### 3.6.4.4 Coherence

Building on metrics like Rényi Entropy, Jensen-Shannon Divergence, and perplexity, we introduce coherence as a crucial metric to assess the semantic interpretability of topics in the hLDA model, ensuring that top words in each topic form meaningful concepts for journal abstract classification (e.g., a "deep learning" topic with words like "neural," "network"). Unlike perplexity, which evaluates predictive accuracy, coherence quantifies how well a topic's top words co-occur in similar contexts within documents, enhancing human-interpretable topic quality [**?** ].

Coherence is computed for each topic node's word distribution $\phi_t = p(w_i \mid t)_{i=1}^{V}$, where $V$ is the vocabulary size, by analyzing the top-$N$ words (e.g., $N = 10$). Two common metrics are Normalized Pointwise Mutual Information and C_V Coherence.

**Normalized Pointwise Mutual Information** (NPMI) measures the co-occurrence strength of word pairs within a topic, indicating how strongly two words are semantically associated within documents of an external corpus (e.g., Wikipedia). For words $w_i$ and $w_j$, NPMI is defined as:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)}$$

where $p(w_i, w_j)$ is the joint probability of the words co-occurring within a document in the corpus, and $p(w_i)$, $p(w_j)$ are their marginal probabilities, estimated from word frequencies within documents. The numerator, $log\frac{p(w_i,w_j)}{p(w_i)p(w_j)}$ is the Pointwise Mutual Information (PMI), which quantifies how much the words co-occur compared to if they were independent.

The denominator $-\log p(w_i, w_j)$ normalizes NPMI to a range of $[-1, 1]$, where,

- NPMI = 1 indicates perfect association (words always co-occur)
- NPMI= 0 indicates independenc
- NPMI= -1 indicates mutual exclusion

NPMI is computed for all pairs of the top-$N$ words in a topic, forming an NPMI matrix $A \in \mathbb{R}^{N\times N}$ with $A_{ii} = 0$ to exclude self-pairs. Higher NPMI scores reflect stronger semantic relationships, making it a foundational measure for topic coherence [23].

C_V Coherence builds on NPMI to assess the overall cohesion of a topic's top-$N$ word set, capturing how aligned the words' co-occurrence patterns are within documents. Using the NPMI matrix $A$, each word's context vector is $v_i = A_{[i:]}$, representing its NPMI scores with all other top-$N$ words. The topic's mean vector is $m = \frac{1}{N}\sum_{i=1}^{N} v_i$, averaging the context vectors. C_V is defined as the average cosine similarity between each word's context vector and the mean vector:

$$\text{C_V} = \frac{1}{N}\sum_{i=1}^{N} \frac{v_i \cdot m}{|v_i||m|}.$$

C_V ranges from 0 to 1, with higher values indicating greater topic cohesion, as the words' co-occurrence patterns align closely with the topic's average pattern. Unlike raw NPMI averages, which focus on pairwise strength, C_V evaluates set-level coherence, remaining robust even sometimes NPMI scores are low (e.g., for near-synonymous words like "neural" and "network" that rarely co-occur within a document but share similar contexts), as their context vectors align directionally [23]. This makes C_V highly correlated with human judgments of topic interpretability.

In later hLDA experiment, if NPMI$(w_i, w_j) < 0$ but C_V $> 0$ for a topic, it indicates that some word pairs within the topic are mutually exclusive (i.e., they rarely co-occur within documents), but their co-occurrence patterns with other words align consistently, contributing to overall topic cohesion [23].

#### 3.6.4.5 Balance degree of tree structure

To evaluate the structural properties of the hierarchical model, we adopt two complementary metrics: **Gini coefficient** of document distribution within a layer and the branching factor of the tree. And **branching factor** between sdjacent layers.

The Gini coefficient, originally introduced to measure inequality in income distribution, is used here to quantify how evenly documents are allocated across nodes within the same layer. Formally, let a layer $l$ contain $n_l$ nodes, each node with $d_i$ documents, and let the total number of documents be $D = \sum_{i=1}^{n_l} d_i$. The document proportion of node $i$ is defined as

$$p_i = \frac{d_i}{D},$$

and the Gini coefficient is computed as

$$\text{Gini} = 1 - \sum_{i=1}^{n_l} p_i^2.$$

A high Gini value indicates that documents are concentrated in a small subset of nodes, while a low value reflects a more balanced allocation.

The branching factor, on the other hand, measures the density of the tree structure by describing the average number of child nodes per non-leaf node. If the tree contains $N$ non-leaf nodes and node $j$ has $c_j$ children, the branching factor is defined as

$$\text{Branch} = \frac{1}{N} \sum_{j=1}^{N} c_j.$$

Together, these two measures provide complementary perspectives on structural quality: the Gini coefficient captures the balance of document assignment across nodes, while the branching factor reflects the granularity of topic subdivision. Structures with moderate inequality and moderate branching are considered most desirable, as they balance interpretability and efficiency while avoiding both overfitting (high Gini with high branching) and underfitting (low Gini with low branching).

## 3.7 Comparative Topic Models

To comprehensively evaluate the performance of our hLDA performance, we compare it against several baseline and alternative topic models, aligning hyperparameters and model scale where possible.

#### 3.7.0.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [4] is the foundational probabilistic topic model, serving as the basis for hierarchical extensions like hLDA. It treats documents as mixtures of topics and topics as multinomial distributions over a vocabulary. Unlike hLDA, LDA is a flat model with all topics on a single level, requiring a predefined number of topics $K$, which contrasts with hLDA's nonparametric approach where topic counts emerge from data. In our comparison, LDA uses collapsed Gibbs sampling for inference, matching hLDA's document-topic $\alpha$ and topic-word $\eta$ priors, with $K$ set to the average number of leaf topics from hLDA.

#### 3.7.0.2 Correlated Topic Model (CTM)

While LDA assumes that, given the document's topic proportions drawn from the Dirichlet prior, the topic assignments of words within a document are independent, it does not explicitly model correlations between topics. For example, a document about "optics" is more likely to also mention "lenses" than "sports". The Correlated Topic Model (CTM) replaces the Dirichlet prior with a logistic–normal distribution, which can capture correlations between topics by introducing a Gaussian prior over topic proportions.

#### 3.7.0.3 Hierarchical Dirichlet Process (HDP)

Hierarchical Dirichlet Process (HDP) extends LDA nonparametrically, allowing the number of flat topics to grow with data, unlike LDA's fixed $K$ or hLDA's hierarchical growth. HDP organizes topics in a single layer using the Chinese Restaurant Franchise (CRF) analogy, where documents (restaurants) share a global menu of topics (dishes) from a Dirichlet Process (DP), with $\gamma$ controlling new topic introduction and $\alpha$ governing per-document topic usage. The base measure $H$ is a Dirichlet prior over words ($\beta$).

In our setup, HDP's concentration parameters are tuned to match hLDA's leaf topic count, isolating the impact of hierarchy. Unlike hLDA's, HDP lacks a tree structure.

# Chapter 4

# Experiments and Results Analysis

## 4.1 Data Preparation

In this study, we first extracted a focused sample of 970 research articles published in *Computational Mechanics* from 2016 through 2022. Computational Mechanics Journal, established in 1986 and published monthly by Springer, consistently ranks among the top journals in computational and applied mechanics. It publishes original research on advanced numerical methods—such as finite element, meshless, and variational techniques—for simulating complex physical systems. Topics span solid and fluid mechanics, fluid–structure interaction, fracture and damage modeling, multi-scale and multi-physics coupling, and emerging computational frameworks for materials and biological systems.

### 4.1.1 Data Cleaning

To prepare the corpus for topic modeling, we conducted a multi-stage data cleaning and normalization process. First, we extracted relevant fields (IDs, abstracts, keywords, years). Each abstract was lowercased and stripped of non-alphabetic characters, tokenized, and stopwords removed. Part-of-speech tagging and lemmatization brought words to root form. To reduce noise, we set a minimum frequency threshold (at least five occurrences across the dataset), keeping only frequent words. Documents were updated to include only these valid tokens. Filtered token lists were rejoined as cleaned texts, with non-essential/outlier records removed. The resulting structured, lemmatized, frequency-filtered corpus was then ready for topic modeling and analysis.

| | |
|---|---|
| Title | Multiscale thermoelastic modelling of composite strands using the "fundamental solutions" method |
| Abstract | A novel computationally effective approach to multiscale thermoelastic modelling of composite structures and its application to a thermomechanical analysis of two ITER superconducting strands is presented. Homogenisation and recovering problems are solved by means of the "fundamental solutions" method, which was expanded to the case of thermoelastic analysis. We describe a general procedure of multiscale analysis on the basis of this method and apply it to recover stresses at the microscopic scale of a composite strands using a two-level procedure. The recovered micro-stresses are found to be in good correlation with the stresses obtained in the reference problem where the entire composite structure was modelled with a fine mesh. |
| **Cleaned Abstract** | novel computationally effective approach multiscale composite structure application thermomechanical analysis two present problem solve mean fundamental solution method case analysis describe general procedure multiscale analysis basis method apply stress microscopic scale composite use procedure find good stress obtain reference problem composite structure model mesh |
| Year | 2022 |

TABLE 4.1: Example entry from the cleaned corpus.

## 4.2 Experimental Setup and Evaluation Protocol

### Convergence Threshold

In Chapter 3, we introduce the R-hat statistic to monitor the convergence of Gibbs sampling across multiple chains for each hyperparameter combination. For each combination of hyper-parameters (*depth*, $\gamma$, $\eta$, $\alpha$), we initialize three independent Markov chains. Convergence is assessed by tracking the stability of stable document counts. When indicators across chains exhibit similar values, convergence is indicated.

Specifically, the R-hat statistic is defined as: $R_{\text{hat}} = \sqrt{\frac{\hat{\text{Var}}}{W}}$ where $\hat{\text{Var}}$ is the estimated variance combining within-chain variance $W$ and between-chain variance $B$. If $R_{\text{hat}}$ value approach 1, an near-perfect convergence is arrived, at that time $B$ and $W$ become nearly equal.

In this study, we adopt a stringent R-hat threshold of 1.1, following established practices , which is more conservative than the 1.2 threshold often deemed sufficient in prior work [1]. This ensures robust convergence assessment for our hLDA model's hyperparameter optimization.

And for the order of Rényi Entropy, we set $q$ =2 to specificlly match the influence caused by *eta*, which will be explained in stage 2 tuning, with a useful learning for the role of eta.

To matigate the fluctuation cased by unstable nodes with little document, who may have large entropy and little JSD, we used the doc size to weighted nodes infulence and we focuse more on stable nodes. For a layer $l$ in the hLDA tree, the document-averaged Rényi Entropy within layer $l$, weighted by document counts, is:

$$\bar{H}_q^{(l)} = \frac{\sum_{k \in D_l} n_k \cdot H_q^{(k)}}{\sum_{k \in D_l} n_k}, \quad H_q = \frac{1}{1-q} \log \left( \sum_{i=1}^{V} p(w_i)^q \right)$$

where $D_l$ denotes the set of nodes at layer $l$, for any node $k$ ind $D_l$, $H_q^{(k)}$ represents the Rényi Entropy of the word distribution for node $k$, and $n_k$ is the number of documents assigned to node $k$. This weighted mean mitigates the influence of nodes with few documents, where the smoothing effect of the Dirichlet prior $\eta$ may lead to more uniform word distributions, inflating the entropy.

We have:

- Back Window =5
- Multiple Chains = 3
- R-hat convergence threshold =1.1
- Sliding window set for calculate R-hat = 20

**Three-Stage Grid Search Framework with Selection Criteria**

Inspired by established practices in topic modeling [1], our three-step grid search optimizes hyper-parameters (*depth*, $\gamma$, $\eta$, $\alpha$) efficiently, minimizing computational load. First, we explore structural parameters *depth* and $\gamma$, fixing $\eta$ and $\alpha$. Next, with the optimal *depth* and $\gamma$, we tune the Dirichlet prior $\eta$, which significantly influences tree structure, followed by $\alpha$, to identify the best configuration for our scenario.

In each stage, we define a candidate list of hyperparameter values and evaluate their performance using specific metrics, including Rényi Entropy and Jensen-Shannon Divergence (JSD) to assess the optimal tree structure, and coherence and perplexity to evaluate semantic representation.

As shown in the table 4.2, our three-step grid search strategy optimizes the hLDA model's hyperparameters systematically. In Stage 1, we explore the structural parameters *depth* and $\gamma$ to establish the initial tree structure. In Stage 2, with the optimal *depth* and $\gamma$ fixed, we tune $\eta$ to refine the topic-word distributions. In Stage 3, we optimize $\alpha$ to balance document-topic assignments, leading to the final optimal combination.

| | **Hyper-parameters** | | | | **Selection Criteria** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Stage | *depth* | $\gamma$ | $\eta$ | $\alpha$ | RE | JSD | NPMI | C_V | Perplexity |
| 1 | √ | √ | | | √ | √ | | | |
| 2 | | | √ | | | | √ | √ | √ |
| 3 | | | | √ | √ | √ | √ | √ | √ |

TABLE 4.2: A series of metrics of structural and distribution prior fields are adopted to evaluate hLDA model performance.

## 4.3 Stage 1: explore tree depth and $\gamma$, fix $\eta = 0.1$, $\alpha = 0.1$

**Testing Set Selecting** For the candidate of depth, we perform a grid search over $depth \in \{3, 4, 5\}$ and $\gamma \in \{0.01, 0.05, 0.1, 0.5, 1\}$ to identify an optimal tree configuration. Totally, we explored 15 combinations of parameters, each combination is evaluated by three independent Markov chains.

We exclude $depth = 2$, as it results in a flat model equivalent to standard LDA, with only a root node and leaf nodes, failing to leverage hLDA's hierarchical advantages. The range $depth \in \{3, 4, 5\}$ is selected based on prior work, such as Blei et al. (2010), which notes that shallow $depth < 3$ cannot effectively capture hierarchical structures, while excessively deep trees $depth > 5$ may lead to sparse sampling and convergence difficulties, particularly with limited document counts (Griffiths et al., 2004).

For $\gamma$, the concentration parameter in the nested Chinese Restaurant Process (nCRP), a higher value increases the probability of creating new paths (Pr(new path) $\propto \gamma$), while a lower value favors reusing existing paths. Following Blei et al. (2010) and Paisley et al. (2015), we test $\gamma \in \{0.01, 0.05, 0.1, 0.5, 1\}$, covering a spectrum from conservative to exploratory path creation. Our analysis indicates that $\gamma = 1$ balances the probability of new path creation with the reuse of existing paths, yielding a sufficient number of paths, and thus serves as an upper threshold. This range aligns with common practices in nonparametric topic models, ensuring a robust exploration of the parameter space.

To isolate the structural effects of *depth* and $\gamma$ in the first stage, we fix the Dirichlet priors $\eta$ and $\alpha$ at balanced values to minimize their influence on tree performance, thereby emphasizing the characteristics of *depth* and $\gamma$.

For $\eta$, the Dirichlet prior for topic-word distributions, a value of $\eta = 0.1$ is selected to smooth the word distribution within topics, enhancing the influence of less frequent words. The posterior probability of generating a word $w_i$ from a topic under this Dirichlet distribution is given by:

$$p(w_i \mid \eta) = \frac{c_{w_i} + \eta}{N + V\eta}, \tag{4.1}$$

where $c_{w_i}$ is the observed frequency of word $w_i$ in the topic, $N$ is the total observed frequency of all words in the topic, and $V\eta$ is the pseudo-count contributed by the prior across the vocabulary of size $V = 1490$. With $\eta = 0.1$, the prior contributes a pseudo-count of $V\eta = 149$, which moderately smooths the distribution without dominating the observed counts. For comparison, $\eta = 1$ yields a pseudo-count of 1490, which would exert a stronger influence. Given that nodes at $depth \in 3, 4, 5$ may contain 30 to 4000 words, setting $\eta = 0.1$ ensures a balanced smoothing effect suitable for the full vocabulary.

For $\alpha$, the Dirichlet prior for document-topic (layer) distributions, we set $\alpha = 0.1$ to apply light smoothing, equivalent to 0.1 pseudo-counts per layer.This value maintains weak sparsity, enabling documents to concentrate primarily on a few layers without excessive bias, particularly given the relatively shallow $depth \in 3, 4, 5$. During the early stages of inference, this prevents zero probabilities and promotes uniform exploration across layers; as document counts accumulate, the data increasingly dominate, ensuring that $\alpha$ does not interfere with the structural evaluation of *depth* and $\gamma$ in Stage 1.

**Analysis After Convergence**

Upon achieving convergence, we evaluate Rényi Entropy and Jensen-Shannon Divergence (JSD) to determine the optimal parameter combination. With $\alpha$ and $\eta$ fixed, we expect that as *depth* increases, word distributions across layers become more granular, leading to a decrease in Rényi Entropy, while JSD increases, reflecting greater topic distinctiveness.

Empirical observation. Across depths $= 3, 4, 5$, the weighted Rényi $H_2$ rises from root toward deeper layers and remains higher at the leaf than at the root, while sibling JSD increases in early layers but drops at the last node. This "high-$H_2$ & low-JSD at the leaf" deviates from the initial expectation that deeper levels should yield more granular (lower-entropy) topics and higher between-topic separation.

These findings deviate from our initial expectations, which were based on the assumption that increasing *depth* would lead to more granular word distributions, reducing Rényi Entropy, and increasing JSD. These findings deviate from our initial expectations and warrant further investigation.

Intuitively and exactly, this result is large effected due to the choice of eta, this shows we choose a larger and not ideal $\eta$ from beginning, we anlysis this result for better

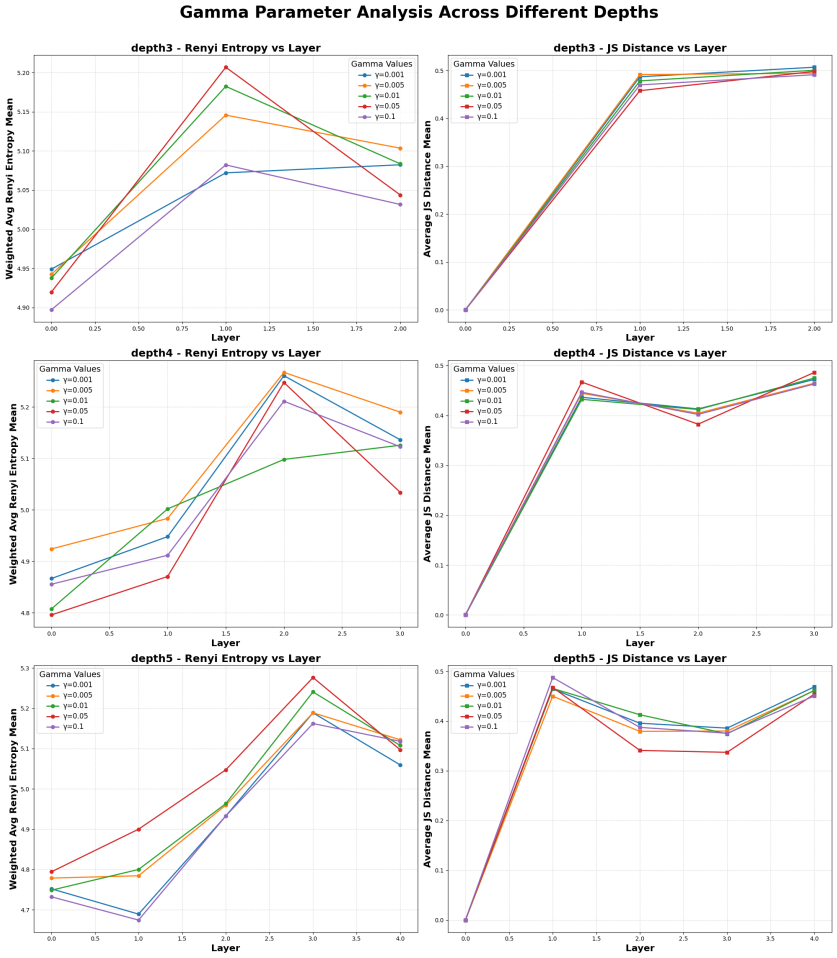


FIGURE 4.1: Analysis of Renyi Entropy and JSD across different model depths and gamma values. The plots on the left show the Weighted Average R

optimization that,

$$p_{w_i}(\eta) = \frac{C_{w_i} + \eta}{N + V\eta} \tag{4.2}$$

$$= \underbrace{\frac{N}{N + V\eta}}_{\lambda} \cdot \underbrace{\frac{C_{w_i}}{N}}_{q_{w_i}} + \underbrace{\frac{V\eta}{N + V\eta}}_{1-\lambda} \cdot \underbrace{\frac{1}{V}}_{u} \tag{4.3}$$

$$= \lambda\, q_{w_i} + (1 - \lambda)\, u \tag{4.4}$$

We have

$$p(\eta) = \lambda\, q + (1 - \lambda)\frac{1}{V}, \quad \lambda = \frac{N}{N + V\eta} \in (0, 1), \quad 1 - \lambda = \frac{V\eta}{N + V\eta}$$

with $q$ the empirical distribution $(\sum_i q_i = 1)$ , $N$ the node token count, and $V$ the vocabulary size. As depth increases (or larger $\gamma$ splits the tree), per-node $N$ decreases so $\lambda$ decreases and the pull toward $u$ strengthens.

For Rényi Entropy ($q$=2), we have

$$H_2(x) = -\log \sum_i (x_i^2)$$

Plug $p(\eta)$ into the definition and expand:

$$\begin{aligned}
H_2\big(p(\eta)\big) &= -\log \sum_i \Big(\lambda q_i + (1-\lambda)\tfrac{1}{V}\Big)^2 \\
&= -\log \sum_i \Big(\lambda^2 q_i^2 + \tfrac{2\lambda(1-\lambda)}{V} q_i + \tfrac{(1-\lambda)^2}{V^2}\Big) \\
&= -\log \Big(\lambda^2 \sum_i q_i^2 + \frac{2\lambda(1-\lambda)}{V} \underbrace{\sum_i q_i}_{=1} + \underbrace{\sum_i \tfrac{(1-\lambda)^2}{V^2}}_{=(1-\lambda)^2/V}\Big) \\
&= -\log \Big(\lambda^2 \sum_i q_i^2 + \frac{1-\lambda^2}{V}\Big) \\
&= -\log \Big(\lambda^2 \Big(\sum_i q_i^2 - \tfrac{1}{V}\Big) + \tfrac{1}{V}\Big).
\end{aligned}$$

Next, we define

$$g(\lambda) = \lambda^2 \Big(\sum_i q_i^2 - \tfrac{1}{V}\Big) + \tfrac{1}{V}, \quad \text{so} \quad H_2\big(p(\eta)\big) = -\log g(\lambda).$$

Then

$$g'(\lambda) = 2\lambda \Big(\sum_i q_i^2 - \tfrac{1}{V}\Big) > 0 \quad \text{whenever } q \neq \text{uniform}.$$

Also

$$\frac{d\lambda}{d\eta} = -\frac{NV}{(N+V\eta)^2} < 0.$$

By the chain rule,

$$\frac{dH_2}{d\eta} = -\frac{1}{g(\lambda)}\, g'(\lambda)\, \frac{d\lambda}{d\eta} \;>\; 0.$$

Hence increasing $\eta$ (or decreasing $N$) strictly increases $H_2$, we arrive the conclusion that

$$\eta \to 0 \ (\lambda \to 1): \quad H_2\big(p(\eta)\big) \to -\log \sum_i q_i^2 = H_2(q),$$

$$\eta \to \infty \ (\lambda \to 0): \quad H_2\big(p(\eta)\big) \to -\log \sum_i (1/V)^2 = \log V.$$

Dirichlet($\eta$) smoothing moves $p(\eta)$ toward the uniform distribution; as $\eta$ increases (or as node token count $N$ decreases), the entropy $H_2$ increases monotonically from $H_2(q)$ toward $\log V$. As depth increases (or $\gamma$ induces more branching), per-node $N$ typically decreases, so $\lambda$ decreases and the pull toward $u$ strengthens.

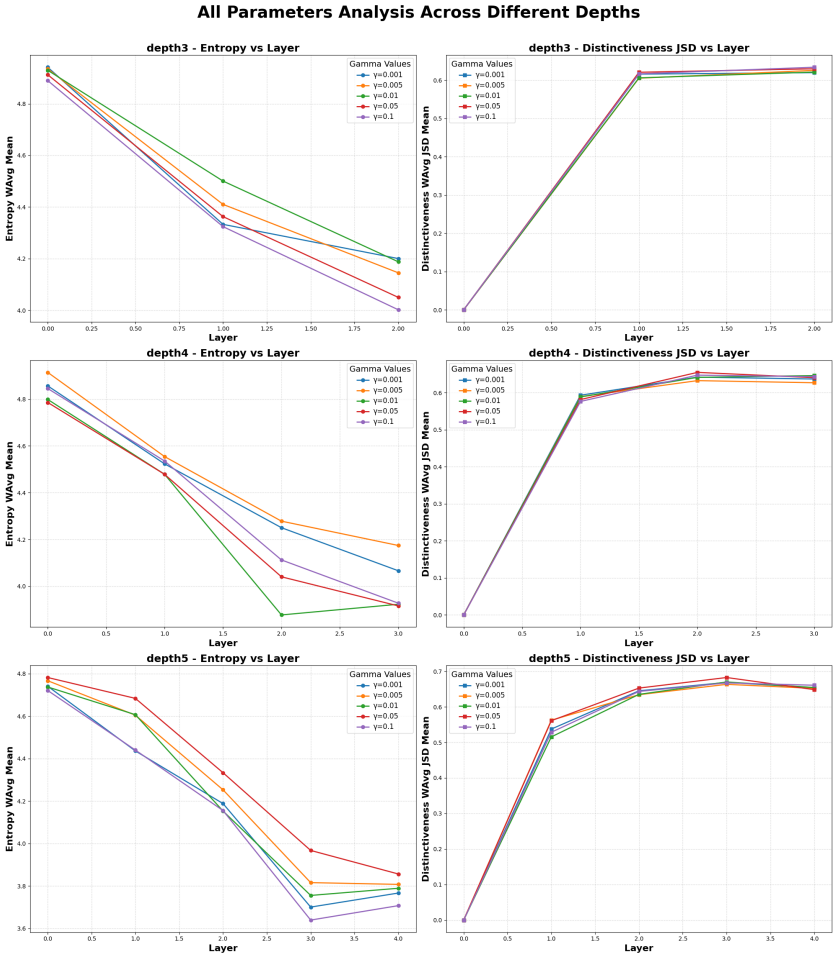


FIGURE 4.2: The plots compare weighted average entropy (left) and distinctiveness JSD (right) across layers for models with depths of 3, 4, and 5 under different gamma ($\gamma$) settings. When depth=3, the consistent decrease in entropy and increase in JSD with layer depth demonstrate that the model effectively learns a hierarchy of increasingly specific and distinct topics

In JSD contraction for two topics $a, b$ and their smoothed versions

$$a_\lambda = (1-\lambda)u + \lambda a, \quad b_\lambda = (1-\lambda)u + \lambda b, \quad m_\lambda = (1-\lambda)u + \lambda \tfrac{a+b}{2},$$

the log–sum inequality (KL's joint convexity) gives

$$\mathrm{KL}(a_\lambda \| m_\lambda) \le \lambda\, \mathrm{KL}(a\|m), \qquad \mathrm{KL}(b_\lambda \| m_\lambda) \le \lambda\, \mathrm{KL}(b\|m),$$

hence

$$\mathrm{JSD}(a_\lambda, b_\lambda) \;\le\; \lambda\, \mathrm{JSD}(a,b)$$

So as $\lambda$ shrinks at deep nodes (smaller $N$), JSD contracts, matching the observed last-layer dip.

**Optimal Parameter** To isolate structural effects, we compute all metrics on the un-

smoothed empirical distributions $q$ (i.e., without applying $\eta$). We evaluate depths 3, 4, and 5 across $\gamma \in 0.001, 0.005, 0.01, 0.05, 0.1$ using the following criteria:

As shown in the figure, the plots reveal distinct patterns: For $depth = 3$, entropy decreases monotonically from root to leaf, and sibling JSD rises steadily without a terminal drop across all $\gamma$ values, with curves tightly clustered, indicating low dispersion.

For $depth = 4$, JSD increases in early layers but exhibits a significant drop at the deepest layer, reflecting the undesired failure mode of reduced topic separability.

For $depth = 5$, JSD plateaus with noticeable variability across $\gamma$, and entropy curves show a wider spread, suggesting over-segmentation and instability.

Based on these observations, we select $depth = 3$ and $\gamma = 0.05$ for Stage I . This choice is justified by $\gamma = 0.05$ yielding the lowest layer-wise dispersion in both weighted average entropy standard deviation and JSD-based distinctiveness standard deviation, while ensuring a steady entropy decline and no JSD collapse at the leaf.

### 4.3.1 Stage 2: Tuning Topic-Word Prior $\eta$

In Stage 1, we gained a valuable insight: $\eta$ significantly influences the overall performance of the topic-word distribution in the hLDA model. A larger $\eta$ smooths the information system, increasing entropy and decreasing Jensen-Shannon Divergence (JSD), resulting in less distinct topics. To address this, we implemented a version without $\eta$ and identified an optimal set of *depth* and $\gamma$ for the next stage.

Notably, when $\eta = 0.1$ in Stage 1, the model's performance was notably reduced, indirectly signaling that adding 149 pseudo-counts (given a vocabulary size $V = 1490$) is excessive for this system, particularly at deeper layers. This observation is supported by the following violin plot, which illustrates token word counts across nodes in each layer. The plot features three violin distributions (from left to right, representing the three layers at $depth = 3$), with black dotted lines indicating the 25%, 50%, and 75% percentiles of word counts of nodes. The root node, with a single value, shows minimal variation, while more than 75% of nodes in Layer 1 and Layer 2 exhibit word counts below approximately 200. This suggests that $\eta = 0.1$, adding 149 pseudo-counts, is unreasonably large, prompting our focus on optimizing $\eta$ in the subsequent stage.

To mitigate the influence of nodes with few documents, where the smoothing effect of the Dirichlet prior $\eta$ results in more uniform word distributions that inflate Rényi Entropy, we compute a document-count-weighted mean to calculate the average Rényi Entropy within each layer.

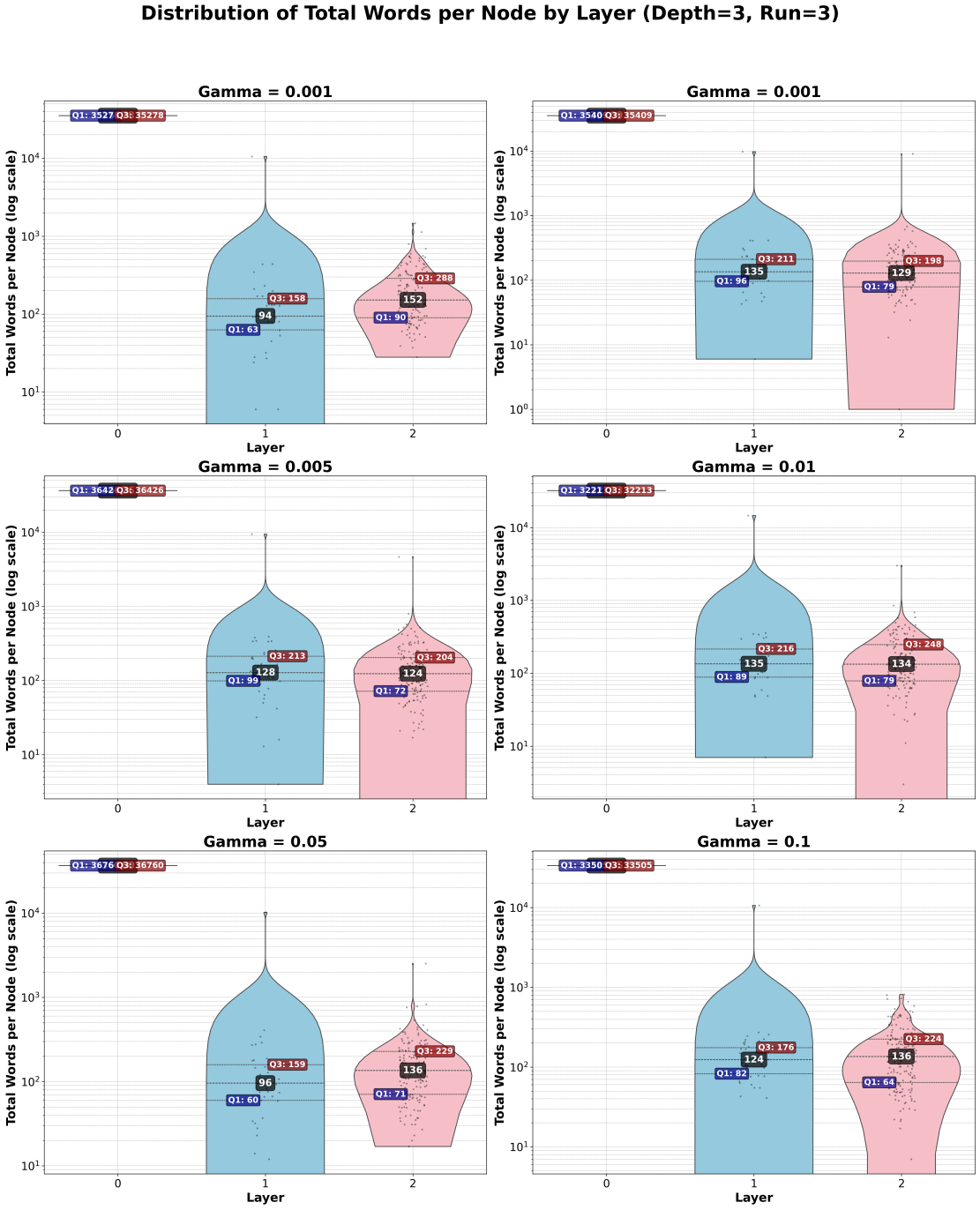


FIGURE 4.3: Violin plots of word counts per node across layers under different Gamma parameters. With the model depth fixed at 3, the figure shows the distribution of total words per node for each layer (Layer 0, 1, 2) as Gamma varies from 0.001 to 0.1. The plots clearly show that the root node (Layer 0) contains the vast majority of words, while the word count distribution (median, quartiles) in deeper layers (Layer 1, 2) adjusts accordingly with the changing Gamma value

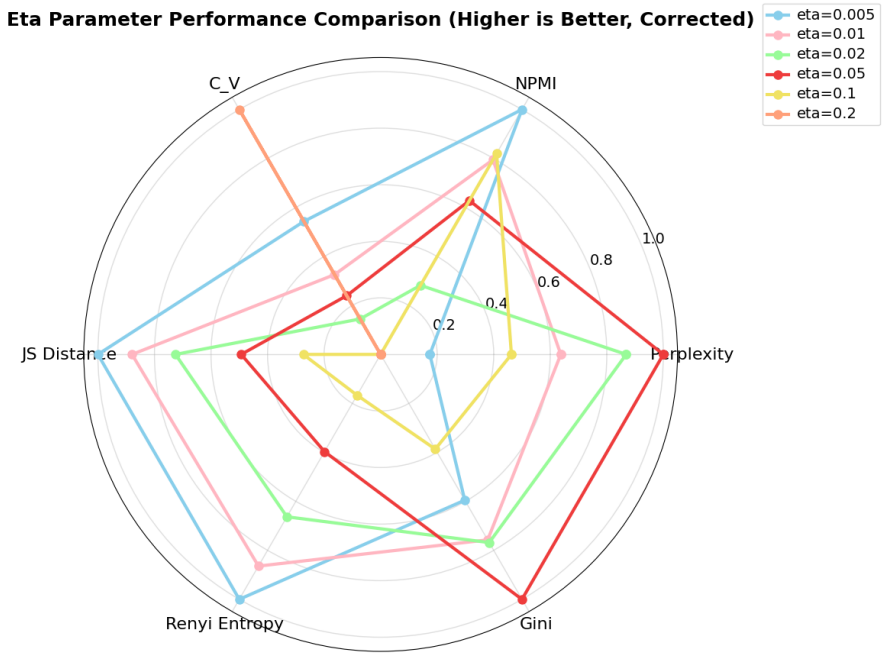


FIGURE 4.4: Radar chart comparing the performance of different Eta parameters (from 0.005 to 0.2) on six key metrics. The chart evaluates Perplexity, NPMI, C_V, JS Distance, Renyi Entropy, and Gini coefficient. All metrics are normalized and unified to a 'higher is better' standard. The results show that eta=0.05 (the red line) exhibits strong overall performance

In Stage 2, we incorporate two metrics, coherence and perplexity, commonly used in topic modeling, to better evaluate the semantic representation of the model. to select the combination that maximises interpretability and modelling accuracy. And we select $\eta \in 0.001, 0.005, 0.01, 0.05, 0.1, 0.2$ to optimize the Dirichlet prior for the hLDA model, driven by the following considerations:

In community Practices and Literature, the default word prior in widely used tools like MALLET is $\beta = 0.01$ [24], a setting adopted by many empirical studies as a baseline [25]. Some studies employ stronger smoothing, such as $\beta = 0.1$, for flatter topic distributions (e.g., Yoshida et al., 2023; Andrzejewski et al., 2009), suggesting that extending the upper bound to 0.1–0.2 is reasonable and conservative [26]. Excessively large priors (e.g., $\eta = 1$ or higher) are used in literature as controls to demonstrate smoothing's impact on perplexity and topic readability, but we exclude these from regular options, using 0.2 as an "over-smoothing" boundary to observe entropy rise and JSD contraction thresholds [27]. Our $\eta$ grid encompasses the commonly used value of 0.01 from tools like MALLET and numerous studies, extends downward to 0.001 and 0.005 for minimal smoothing, includes the effective range of 0.05–0.1, and reaches 0.2 to detect over-smoothing effects. This design aligns with MALLET defaults, literature benchmarks, and the logarithmic search convention for hyperparameters.

In this study, we transform six metrics into a radar chart to evaluate the performance of our hLDA model. Notably, despite Rényi Entropy and Perplexity being naturally "lower is better" metrics, in order to ensure all metrics adhere to the "higher is better" principle, we apply a two-step normalization process to the hLDA model evaluation data. First, we perform Min-Max scaling on all metrics, scaling their values to the [0, 1] interval using the formula:

**Min-Max Scaling**
Each metric is first scaled to the [0, 1] interval using the Min-Max formula:

$$\text{scaled_value} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

where $x$ is the current metric value, $x_{\text{min}}$ is the minimum value across all observations, and $x_{\text{max}}$ is the maximum value. This initial step standardizes the range, with the distinction handled in the subsequent adjustment.

**Direction Alignment** For metrics where "higher is better" (JSD, C_V, NPMI), the normalized value remains the scaled value:

$$\text{normalized_value} = \text{scaled_value}$$

For metrics where "lower is better" (Perplexity, Gini, Rényi Entropy), the scaled value is inverted:

$$\text{normalized_value} = 1 - \text{scaled_value}$$

This two-step process ensures all metrics are unified on the [0, 1] scale, where higher values consistently indicate better performance.

We begin by examining the radar chart in Figure to assess the hLDA model's performance across six normalized metrics with $depth = 3$ and $\gamma = 0.05$ fixed. The chart reveals among six $\eta$, when $\eta = 0.05$ (red line) achieves a balanced performance that:

1. near-optimal scores for Perplexity =356.50 (No.1) , Gini = 0.3857 and Rényi Entropy = 4.4826
2. moderate scores for C_V = 0.5701, NPMI = 0.0766

As visualized in the figure, JSD and Rényi Entropy notably increase as $\eta$ decreases, a trend driven by reduced smoothing effects. This allows the raw data (word frequencies) to dominate, yielding sharper, more distinct topics, which lowers actual entropy and boosts JSD. As a result, these metrics offer limited reference value due to their sensitivity to smoothing adjustments.

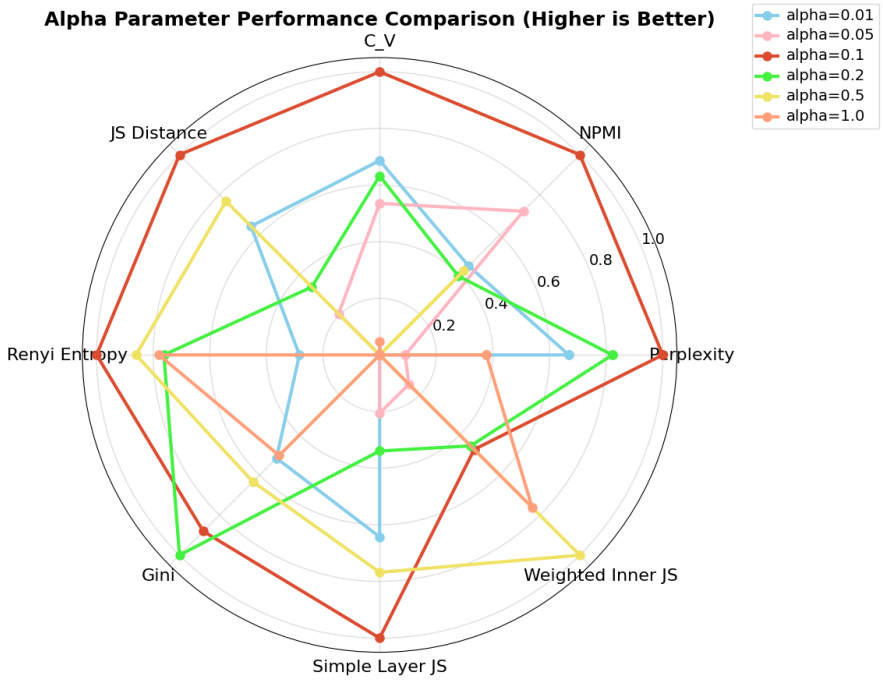


FIGURE 4.5: Radar chart comparing the performance of different alpha parameter values across eight evaluation metrics. All metrics are normalized so that a higher value (closer to the outer edge) indicates better performance. The analysis shows that an alpha value of 0.1 (red line) provides the most robust and balanced performance, achieving high scores on topic coherence (NPMI, C_V), topic distinctiveness (Weighted JSD, Simple Layer JS), and perplexity.

Based on this analysis, we select $\eta = 0.05$ as the optimal Dirichlet prior for topic-word distributions, offering superior model fit, structural balance, and semantic quality. This choice is finalized with

$$depth = 3, \gamma = 0.05, \eta = 0.05$$

### 4.3.2 Stage 3: Optimizing Topic Diversity with $\alpha$

With $depth = 3$, $\gamma = 0.05$, and $\eta = 0.05$ fixed, we meticulously tune the document-topic concentration parameter $\alpha$ across a log-spaced grid $0.01, 0.05, 0.1, 0.2, 0.5, 1$. The range of $\alpha$ spans $10^{-2}$ to 1 are typical values used in LDA/hLDA studies, as evidenced by Griffiths & Steyvers (2004) [17], who employed $\alpha = 50/T$ (yielding 0.5 to 1 for $T \in [50, 100]$), and MALLET's default optimization within this range [24]. And topic models package also adopts similar defaults for Gibbs and Variational EM routines, ensuring comparability. Furthermore, prior research underscores the critical role of $\alpha$, a hyperparameter smoothing significantly influences both performance and algorithmic behavior (Wallach et al., 2009; Asuncion et al., 2009) [28], necessitating a broad exploration across orders of magnitude. This strategic selection of the $\alpha$ grid provides a robust foundation for our optimization process.

As shown in the figure, the trends across these $\alpha$ values reveal distinct performance patterns and $\alpha = 0.1$ emerges as the optimal choice achieving the lowest perplexity, the highest Normalized Pointwise Mutual Information (NPMI), and the second-highest C_V coherence, while maintaining moderate Gini coefficients and branching factors. Additionally, Jensen-Shannon Divergence (JSD) exhibits a consistent increase, peaking at its largest value with $\alpha = 0.1$, and Rényi Entropy shows a consistent decrease, indicating improved topic granularity. These trends suggest that $\alpha = 0.1$ strikes an optimal balance between semantic coherence, structural stability, and predictive accuracy. Consequently, we select $\alpha = 0.1$ as the preferred setting for Stage 3, finalizing the hyperparameter

$$\alpha = 0.1, \quad \text{with } depth = 3, \quad \gamma = 0.05, \quad \eta = 0.05$$

## 4.4 Model Comparation

In our pursuit to unravel the distinctive strengths of the hLDA model within the realm of topic modeling, we've chosen three trusted companions: Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM), and Hierarchical Dirichlet Process (HDP), each offering a distinct perspective to illuminate hLDA's path.

The story begins with **LDA**, the foundational storyteller of flat, parametric topics, a direct ancestor to hLDA that sets a simple stage by fixing the number of topics $K$. This allows us to see how hLDA's nonparametric, hierarchical extension, built on the nested Chinese Restaurant Process (nCRP), enhances interpretability by adding depth where LDA remains flat. After optimizing hyperparameters through three simulation runs, we determined that Layer 1 averages 59 nodes, while Layer 2 (the leaf layer) comprises 252 nodes. To align the modeling settings between hLDA and LDA—aside from the hierarchical depth—we adopt the same $\alpha = 0.1$, $\gamma = 0.05$, and $\eta = 0.05$ parameters in LDA, treating its $K = 252$ topics as equivalent to hLDA's leaf layer for a fair comparison.

Next, we encounter **CTM**, a sophisticated cousin that enriches the narrative by modeling topic correlations through a logistic-normal prior, stepping away from LDA's assumption of topic independence. Unlike hLDA's layered structure, CTM operates on a flat plane, capturing relationships between topics to improve semantic coherence without hierarchical depth. To ensure consistency, we configure CTM with the same number of topics ($K = 252$) and hyperparameters ($\alpha = 0.1$, $\eta = 0.05$), adjusting only its correlation parameter to reflect its unique focus, allowing us to contrast hLDA's hierarchical gains with CTM's correlation-driven insights. This comparison reveals whether hLDA's enhanced interpretability stems from its depth or CTM's inter-topic connections.

| Model | *depth* | $\gamma$ | $\eta$ | $\alpha$ | Leaf nodes | Branching factor |
|---|---|---|---|---|---|---|
| hLDA | 3 | 0.05 | 0.05 | 0.10 | 252 | L159, L24 |
| PAM | 3* | — | 0.05 | 0.10 | k2=4 (sub layer) | k1=59 (super layer) |
| LDA | — | — | 0.05 | 0.10 | k=252 | — |
| CTM | — | — | 0.05 | 0.10* | k=252 | — |

Finally, **HDP** joins as a nonparametric ally, dynamically inferring topic numbers through the stick-breaking process, a mechanism that hinges on tuning the concentration parameter to shape the number of nodes. Unlike hLDA's tree-like hierarchy enabled by nCRP, HDP maintains a flat structure, and we test various concentration values—starting with a base aligned to $\alpha = 0.1$ and $\eta = 0.05$—to match hLDA's 252 leaf nodes, omitting $\gamma$ due to its lack of branching. This adjustment lets us explore whether hLDA's advantages arise from its nonparametric flexibility or its specific hierarchical design.

Together, LDA (flat and parametric), CTM (flat with correlations), and HDP (nonparametric and flat) span the diverse landscape of topic model designs, providing a robust benchmark to isolate hLDA's hierarchical contributions

This table summarizes the parameter configurations and structural characteristics of hLDA and its baselines, ensuring alignment for comparative analysis.

It is worth noting that, for CTM, we initially considered setting the number of topics equal to the average number of hLDA leaf nodes (≈252). However, the computational complexity of CTM is substantially higher than LDA or HDP, since it requires estimating a full covariance matrix $\Sigma \in \mathbb{R}^{K \times K}$. The number of free parameters grows quadratically with $K$: at $K = 252$, this involves over 63,000 covariance terms, which significantly increases both memory consumption and training time.

Even after upgrading our hardware environment to RTX 5090 GPU with 24 CPU cores, training CTM with $K = 252$ failed to converge within reasonable time. As a practical compromise, we reduced the number of topics to $K = 128$, approximately half the number of hLDA leaves. This choice lowers the covariance parameter space to about 16,000 terms, making training computationally feasible, while still providing sufficient topic granularity for comparison. Moreover, selecting 128—being a power of two—offers additional efficiency benefits in matrix computations. And We observe that Perplexity in CTM model will decrease with $K$ then increase, and Perplexity($K$=100) ¡ Perplexity($K$=128), leading to Perplexity(hLDA) ¡ ¡ Perplexity($K$=252).

**Covergence**: The table indicates that LDA and HDP converge quickly with stable log-likelihood values across runs, however, CTM struggles with slower convergence and higher variance, a trade-off from its logistic-normal prior's complexity. hLDA exhibits moderate stability, with convergence influenced by $\gamma$ and *depth*, reflecting its intricate

| **Model** | **Perplexity** | Coherence (C_V) | Coherence (NPMI) | Topic Count |
|---|---|---|---|---|
| LDA | 750 | $0.48 \sim 0.50$ | $-0.20$ | 252 (fixed) |
| HDP | 748 | $0.44 \sim 0.45$ | $-0.24$ | 253(dynamic) |
| CTM | 4400–4600 | $0.63 \sim 0.66$ | $0.07 \sim 0.09$ | 128 (fixed) |
| hLDA | 861.8533 | 0.57 | 0.08 | 252 (leaf) |

hierarchical design. This adaptability, while less stable than flat models, enables hLDA to tailor its structure to the data, a key advantage in complex applications.

**Predictive Performance**: LDA and HDP's Perplexity edge with values around 750, excelling in word-level prediction. hLDA follows with a moderate 861.9, balancing predictive power with hierarchical depth, while CTM's high range (4400–4600) underscores its weakness, burdened by correlation modeling.

**Topic Coherence**: C_V and NPMI offer critical insights. LDA and HDP show negative NPMI (-0.20 and -0.24), indicating weak semantic ties despite decent C_V (0.48$\sim$ 0.50 and 0.44$\sim$0.45). As we discussed in Method that when C_V is large and NPMI is small, the words are excludes within nodes. hLDA, with highest NPMI at 0.08 and second high C_V at 0.57, achieves a balanced coherence, leveraging its hierarchy to forge meaningful topic relationships, outshining flat models in structured interpretability. CTM leads with high C_V (0.63–0.66) and positive NPMI (0.07–0.09), driven by correlation modeling.

In conclusion, hLDAstands out for its combination of balanced consistency and hierarchical interpretability, leveraging its layered structure to refine topic relationships. LDA and HDP, as flat models, excel in predictive consistency but lack the depth to organize topics hierarchically, limiting their semantic richness. CTM, while strong in capturing topic correlations, suffers from high perplexity, undermining its predictive performance compared to hLDA's insights from hierarchical structured topics.

## 4.5 Experiment Result Analysis

Following the convergence of three chains under our optimal hyperparameter settings ($depth = 3$, $\alpha = 0.1$, $\eta = 0.05$, $\gamma = 0.05$), we conducted a detailed analysis of the resulting tree structure's stability. This includes assessing:

- the stable document assignments to the same path over the last five iterations
- the persistence of nodes across these iterations
- the consistency of the top words within stable nodes

The three observations collectively reflecting the stabilization of document-to-path assignments and word-to-node allocations.

After approximately 800 iterations, three chains converged with an R-hat value of 1.1, prompting us to select one chain for further examination. For chain 1, 589 papers consistently selected the same path over five iterations. At the final iteration, Layer 1 comprised 63 nodes, all of which remained stable over the past five iterations, while Layer 2 contained 251 nodes, with 185 exhibiting stability. Among these stable nodes, 158 displayed fixed top words, effectively defining their semantic clusters.

To visualize this stability, we extracted a subtree depicted in the accompanying figure, where green nodes represent stable entities persisting across the last five iterations, and orange nodes indicate those that disappeared at least once.

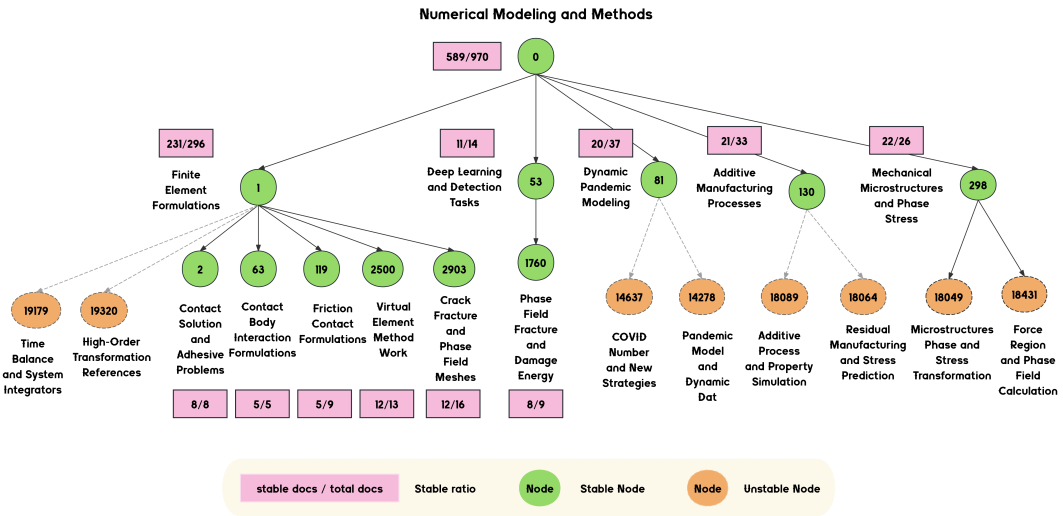


FIGURE 4.6: Visualization of a sub-tree from the hLDA-generated hierarchy, optimized with parameters (depth = 3, $\gamma = 0.05$, $\eta = 0.05$, $\alpha = 0.1$). Green nodes indicate stable nodes consistent across the past 5 rounds, while orange nodes represent unstable ones. Derived from 970 *Computational Mechanics* abstracts (2016–2022), it reflects domains from "numerical modeling" to specialized topics like "finite element formulation." The pink box highlights the stable-to-total document ratio in persisted nodes.

Leveraging online resources, we assigned semantic domains to each node based on its top words with principle that, stable nodes prioritized consistent words, while unstable nodes (existing but with shifting documents) relied on variable words. The pink box in the figure highlights the ratio of stable documents to total documents (including unstable ones) in persisted nodes . This sub-tree is depicted from our generated route-domain Excel file detailing topics' domain along routes, accompanied by the top 10 words for each node.

Given our dataset focuses on computational mechanics, the derived domains naturally align with this field, progressing from broad concepts like "numerical modeling and methods" at the root to specific subdomains. For instance, Node 1, a clustering su node, is labeled "finite element formulation"—a widely recognized method in mechanics calculations, while its leaf nodes further specialize into "Contact Solution and Adhesive Problems," "Contact Body Interaction Formulations," "Crack Fracture and Phase Field Meshes," and other similar scenarios, all leveraging finite element methods. This categorization proves highly human-readable, mirroring the natural organization of computational mechanics literature, where general principles evolve into specialized applications.

Moreover, the hierarchical structure's consistency with this domain is reinforced by leaf nodes that often cluster documents with high internal semantic coherence, as validated by cosine similarity and document stability metrics from prior sections. This capability enables effective classification and retrieval of unlabeled data, fulfilling our initial expectations. We have reached the endpoint we envisioned from the outset.hLDA's ability to construct a stable and semantically meaningful hierarchy underscores its effectiveness, making it a particularly suitable tool for analyzing complex, layered datasets such as our scientific corpus.

# Chapter 5

# Discussion

The experiments conducted in this study prove the efficiency of the proposed three-stage hyperparameter search strategy in optimizing the hierarchical Latent Dirichlet Allocation (hLDA) model for semantically dense corpora, such as the 970 research abstracts from *Computational Mechanics* (2016-2022). By systematically tuning structural and distributional parameters, the approach achieves a robust balance between semantic coherence, structural stability, and predictive accuracy.

In Stage 1, the selection of depth = 3 and $\gamma = 0.05$ proved optimal, as evidenced by decreasing weighted Rényi Entropy (H2) and lower layer-wise dispersion in JensenShannon Divergence (JSD), alongside a steady entropy decline and no terminal JSD collapse at the leaf layer. These findings align with prior work suggesting that shallow hierarchies (depth $< 4$) better capture meaningful structures in limited corpora, while excessive depth leads to sparse sampling and instability [17**?** ]. The choice of $\gamma = 0.05$ further balanced the creation of new paths with the reuse of existing ones, ensuring a stable tree with an average branching factor of 4.31 and 252 leaf nodes, reflecting the nuanced hierarchical relationships within computational mechanics.

Stage 2 revealed the critical role of the topic-word Dirichlet prior $\eta$ in shaping topic distinctiveness. The analysis showed that $\eta = 0.1$, initially used in Stage 1, introduced excessive smoothing (149 pseudo-counts for a vocabulary of 1490), inflating entropy and reducing JSD, which led to less distinct topics at deeper layers. By contrast, $\eta = 0.05$ achieved a superior compromise, yielding the lowest perplexity (356.50 at the leaf layer), a balanced Gini coefficient (0.3857), and competitive coherence (C_V = 0.5701, NPMI = 0.0766). This underscores the importance of minimal smoothing to preserve the empirical word distribution, particularly for nodes with low document counts, where Dirichlet smoothing can partially inflate entropy. Stage 3 finalized the optimization by selecting $\alpha = 0.1$, which regulated document-path sparsity without compromising

semantic quality. This configuration produced the highest NPMI, second-highest $C_v$, and a consistent JSD increase, indicating improved topic granularity and stability. The radar chart visualization, with normalized metrics aligned to a "higher is better" scale, confirmed $\eta = 0.05$ and $\alpha = 0.1$ as a balanced configuration, outperforming both lower ($\eta = 0.005, 0.01$) and higher ($\eta = 0.1, 0.2$) values across perplexity, coherence, and structural metrics.

A key contribution of this study lies in its novel convergence detected framework, which addresses the limitations of traditional log-likelihood based methods often inadequate for nonparametric models like hLDA. By integrating the R-hat statistic (threshold = 1.1) across three parallel Markov chains, document stability counts, and Jaccard similarity of document-path assignments, the framework provides a robust, multi-faceted signal of convergence. This approach mitigates the risk of premature termination or false stability, ensuring that the inferred hierarchy aligns with the instinct nature under the same hyper-parameters setting. The use of weighted Rényi Entropy, weighted by document counts, further counters the smoothing-induced bias of low-document nodes, enhancing the reliability of structural evaluation. Compared to prior work relying solely on perplexity or coherence [4**?** ], this triangulated framework offers a more nuanced understanding of convergence, particularly for hierarchical models where structural integrity is as critical as semantic quality.

When benchmarked against related models, hLDA demonstrates distinct advantages for hierarchical interpretability. Flat models like Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) achieved comparable perplexity ( 750 and 748, respectively) but lacked the nested structure critical for capturing parent-child relationships in scientific abstracts. Their single-layer topic assignments resulted in less granular topics, limiting their ability to reflect the multi-level taxonomies inherent in computational mechanics. The Correlated Topic Model (CTM), with its logistic-normal prior, excelled in semantic coherence ($C_v \approx$ 0.63-0.66, NPMI = 0.07-0.09) by modeling topic correlations but incurred significantly higher perplexity ( 4400-4600), indicating poor predictive performance. In contrast, hLDA's 252 leaf topics, derived through the Nested Chinese Restaurant Process (nCRP), balanced interpretability and predictive accuracy, leveraging hierarchical structures to enhance semantic robustness. This aligns with findings that nonparametric priors like nCRP are particularly suited for complex, label-free corpora [3].

Despite these strengths, the study has limitations that further consideration. The experiments were conducted on a single-domain corpus (*Computational Mechanics*), hLDA's reliance on a single-parent node structure, enforced by the Nested Chinese Restaurant Process (nCRP), may struggle to model cross-disciplinary topics that span multiple

parent categories, such as bioinformatics or sociology modeling. Additionally, the three-stage grid search, though more efficient than exhaustive methods, remains computationally intensive due to the multi-chain Gibbs sampling required for hLDA's inference. The reliance on metrics like Rényi Entropy and JSD, while effective, also requires further validation across other hierarchical Bayesian models to establish their universality. Finally, the framework's focus on unsupervised settings limits its applicability to scenarios with partial labels, where semi-supervised variants like HSLDA or hLLDA might offer additional insights [**?** ].

These limitations pave the way for future research, as outlined in Chapter 6. Extending the framework to multi-domain corpora could enhance its generalizability, potentially by incorporating domain-adaptive priors or transfer learning techniques. Integrating constrained word priors, such as those based on ontologies like WordNet, could further refine topic granularity. Additionally, exploring neural variational inference offers a promising avenue for scaling hLDA to larger datasets, reducing the computational burden of Gibbs sampling. By addressing these challenges, this study contributes a robust foundation for hierarchical topic modeling, offering practical tools for automated knowledge discovery in complex, unstructured texts.

# Chapter 6

# Conclusion and future work

This study has been a thrilling journey into the knowledge of Hierarchical Latent Dirichlet Allocation (hLDA), unlocking the hidden structures within 970 *Computational Mechanics* abstracts (2016–2022). By weaving together Bayesian nonparametric principles with a novel three-stage hyperparameter search, we've crafted a robust framework that brings order to the chaos of unstructured scientific texts.

The three-stage hyperparameter search strategy for hLDA optimizes model performance efficiently. Stage 1 sets a stable hierarchical structure (depth = 3, $\gamma = 0.05$), aligning with the corpus's natural hierarchy. Stage 2 tunes $\eta = 0.05$, ensuring clear, coherent topics ($C_v = 0.5701$, NPMI = 0.0766). Stage 3 optimizes $\alpha = 0.1$, balancing document–path assignments for interpretability. This sequential approach minimizes computational waste and enhances topic tree interpretability. It excels in building data-driven knowledge frameworks for domains like computational mechanics.

The real star of this work is our convergence monitoring framework, a fresh take on taming the unpredictability of nonparametric models. By blending stable document counts, R-hat (threshold = 1.1) across three Markov chains, and Jaccard similarity, we've ensured hLDA's inference is both statistically sound and semantically meaningful, leaving traditional log-likelihood methods in the dust [12]. When pitted against baselines, hLDA shines brightly. Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) matched our perplexity ( 750 and  748) but couldn't capture the nested beauty of scientific taxonomies. The Correlated Topic Model (CTM) dazzled with higher coherence ($C_v \approx 0.63$–$0.66$) but stumbled with sky-high perplexity ( 4400-4600). hLDA, with its Nested Chinese Restaurant Process (nCRP), strikes a delightful balance, making it a natural fit for label-free, complex corpora [3]. This work lights the way for automated knowledge discovery, turning dense abstracts into clear, hierarchical insights.

This study's success with *Computational Mechanics* abstracts lays a solid foundation, but there's exciting work ahead to make hLDA even more powerful. Our experiments thrived in a single-domain corpus, but fields like biology or social sciences, with their rich interdisciplinary topics like bioinformatics, pose a challenge for hLDA's single-parent structure, driven by the Nested Chinese Restaurant Process (nCRP). A key next step is to compare hLDA with multi-parent models like the Pachinko Allocation Model (PAM) on diverse corpora, aiming to better capture cross-disciplinary relationships and strengthen generalizability .

The computational demands of multi-chain Gibbs sampling also call for innovation. Exploring neural variational inference (NVI) could streamline hLDA for larger datasets, maintaining its semantic clarity while boosting efficiency [29]. Another promising direction is to integrate constrained word priors, using tools like frequent pattern mining (FP-tree) or ontologies like WordNet, to refine topic granularity and align with domain knowledge, addressing entropy inflation in low-document nodes as seen in Stage 2. To ensure our evaluation metrics—Rényi Entropy, Jensen–Shannon Divergence (JSD), and R-hat—are robust across contexts, testing them on other hierarchical models like HDP or HSLDA is essential [5]. Finally, deeper exploration of $\alpha$ and $\eta$ interactions across varied corpora could lead to adaptive tuning strategies, making hLDA more flexible for different domains. These steps will push hierarchical topic modeling forward, unlocking new possibilities for structuring complex, interdisciplinary texts with precision and impact.

# Bibliography


[1] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010. doi: 10.1145/1667053.1667056.

[2] Carlos N Silla Jr and Alex A Freitas. A survey of hierarchical classification across different application domains. In *Data Mining and Knowledge Discovery*, volume 22, pages 31–72, 2011.

[3] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84, 2012.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Adler J Perotte, Frank Wood, Noémie Elhadad, and Nicholas Bartlett. Hierarchically supervised latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[6] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, volume 1, pages 495–501, 2000.

[7] Jun Seok Kang, Seongmin Park, and Suma Chari. Hetero-labeled lda: A partially supervised topic model with heterogeneous label types. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 835–845, 2014.

[8] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22:288–296, 2009.

[9] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. *Human Language Technologies: NAACL*, pages 100–108, 2010.

[10] Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Incremental algorithms for hierarchical classification. In *Journal of Machine Learning Research*, volume 7, pages 31–54, 2006.

[11] Sergei Koltcov, Vera Ignatenko, Zeyd Boukhers, and Steffen Staab. Analyzing the influence of hyper-parameters and regularizers of topic modeling in terms of renyi entropy. *Entropy*, 22(4):394, 2020.

[12] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[13] Christian P Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[14] Luke Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1701–1728, 1994.

[15] Daniel Maier, Matthias F. Waldherr, and Markus E. Leiner. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Journal of Communication*, 68(2):254–262, 2018.

[16] Roman Egger and Joanne Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7:886498, 2022. doi: 10.3389/fsoc.2022.886498.

[17] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235, 2004. doi: 10.1073/pnas.0307752101.

[18] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 577–584. ACM, 2006. doi: 10.1145/1143844.1143917.

[19] David Mimno, Wei Li, and Andrew McCallum. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 545–552. ACM, 2007.

[20] David M. Blei and John D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007. doi: 10.1214/07-AOAS114.

[21] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[22] Sonia Jain and Radford M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

[23] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408. ACM, 2015. doi: 10.1145/2684822.2685324.

[24] R Core Team. Mallet: Machine learning for language toolkit. https://cran.r-project.org/web/packages/mallet/index.html. R package documentation.

[25] Proceedings of machine learning research. Includes Tang et al. (2014) and Andrzejewski et al. (2009).

[26] Various. Pubmed central (pmc), . Includes Yoshida et al. (2023).

[27] Various. arxiv preprint server, . Includes over-smoothing studies.

[28] Hanna M. Wallach, David M. Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems 22 (NeurIPS)*, pages 1973–1981, 2009.

[29] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. *Proceedings of the 34th International Conference on Machine Learning*, 48:2410–2419, 2016.