

## CẢI TIẾN THUẬT TOÁN TỐI ƯU GIẢI BÀI TOÁN SUY DIỄN HẬU NGHIỆM VỚI MÔ HÌNH CHỦ ĐỀ

Dương Thị Nhung, Bùi Thị Thanh Xuân\*

Trường Đại học Công nghệ thông tin và truyền thông - ĐH Thái Nguyên

### TÓM TẮT

Bài toán suy diễn hậu nghiệm cho mỗi văn bản đóng vai trò quan trọng trong mô hình chủ đề. Tuy nhiên, trong quá trình giải bài toán suy diễn này thường đưa về dưới dạng một bài toán tối ưu không lồi với dữ liệu lớn, do đó nó thường là bài toán NP-khó. Có nhiều phương pháp được đề xuất để giải xấp xỉ bài toán suy diễn hậu nghiệm như phương pháp Variational Bayes (VB), collapsed variational Bayes (CVB) hay phương pháp collapsed Gibbs sampling (CGS),... Tuy nhiên các phương pháp này hầu hết không đảm bảo về chất lượng cũng như tốc độ hội tụ của thuật toán. Với ý tưởng sử dụng thuật toán Online Frank-Wolfe (OFW) và thuật toán Online Maximum a Posterior Estimation (OPE), chúng tôi đề xuất hai thuật toán cải tiến có hiệu quả giải bài toán suy diễn hậu nghiệm với mô hình chủ đề, đó là IOPE1, IOPE2. Bằng việc sử dụng biên ngẫu nhiên, xấp xỉ ngẫu nhiên và phân phối ngẫu nhiên như phân phối Uniform, phân phối Bernoulli, các đề xuất của chúng tôi được sử dụng để phát triển các phương pháp mới có hiệu quả để học các mô hình chủ đề từ bộ sưu tập văn bản lớn. Các kết quả thực nghiệm cho thấy các phương pháp tiếp cận của chúng tôi thường hiệu quả hơn các phương pháp trước đó.

**Từ khóa:** Suy diễn hậu nghiệm, OPE, Online Frank-Wolfe, mô hình chủ đề, học trực tuyến, xấp xỉ ngẫu nhiên.

*Ngày nhận bài: 11/3/2019; Ngày hoàn thiện: 03/4/2019; Ngày duyệt đăng: 07/5/2019*

## IMPROVEMENT OPTIMIZATION ALGORITHMS APPLIED FOR SOLVING THE POSTERIOR INFERENCE PROBLEM IN TOPIC MODELS

Duong Thi Nhung, Bui Thi Thanh Xuan\*

University of Information and Communication Technology - TNU

### ABSTRACT

The posterior inference problem for individual text plays an important role in the topic models. However, in solving this problem, it is usually given as a nonconvex optimization problem with the large datasets, so it is often NP-hard. There are many methods proposed to approximate the posterior inference problem such as Variational Bayes (VB), collapsed variational Bayes (CVB) or collapsed Gibbs sampling (CGS) methods, but these methods do not guarantee the quality or convergence rate. Using the idea of Online Frank-Wolfe algorithm (OFW) and Online Maximum a Posteriori Estimation (OPE) algorithm, we propose two efficient algorithms for solving the posterior inference problem in the topic models which are IOPE1 and IOPE2. Using stochastic bounds, stochastic approximation and probability distributions such as uniform distribution, Bernoulli distribution, our improvements are used to develop new effective method for learning LDA from large text collections. Experimental results show that our approaches are often more effective than OPE.

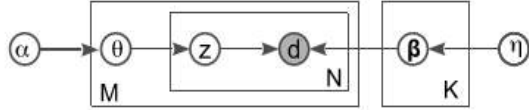
**Keywords:** OPE, Online Frank-Wolfe, topic models, online learning, stochastic approximation, nonconvex optimization.

*Received: 11/3/2019; Revised: 03/4/2019; Approved: 07/5/2019*

\* Corresponding author: Tel: 0902 001581; Email: bttxuan@ictu.edu.vn

**1. Mô hình chủ đề và bài toán suy diễn hậu nghiệm**

Trong mô hình chủ đề Latent Dirichlet Allocation (LDA) [1], mỗi văn bản được biểu diễn theo dạng “bag of word”, tức là mỗi văn bản coi như một túi từ, thứ tự các từ trong văn bản không quan trọng.



**Hình 1.** Mô hình Latent Dirichlet Allocation

Tác giả Blei và các cộng sự [1] đưa ra giả thuyết về cấu trúc ẩn chứa trong tập các văn bản. Mỗi văn bản là sự trộn lẫn của các chủ đề ẩn (latent topics), trong đó mỗi chủ đề là một phân phối của tất cả các từ trong tập từ điển. Mỗi văn bản trong tập corpus được xem như một túi các từ, các từ sinh ra là tổ hợp của các chủ đề mà tác giả muốn viết. Mỗi chủ đề là phân phối của các từ trong tập từ điển. Mô hình sinh được mô tả như sau:

- Với mỗi topic trong tập  $\{1,2 \dots K\}$
- Lấy mẫu  $\beta_k \sim Dir(\eta)$

Sinh văn bản  $w$  có độ dài  $N$ :

- Lấy mẫu  $\theta \sim Dir(\alpha)$
- Với mỗi từ  $w_n$  trong  $N$  từ:
  - + Chọn một topic  $z_n \sim Multinomial(\theta)$
  - + Chọn một từ  $w_n$  với xác suất  $p(w_n | \beta_{z_n})$

Trong các bước học tham số trong mô hình LDA, có hai bước chính là:

- Bước E: suy diễn vector tỉ lệ chủ đề  $\theta_d$  cho từng văn bản  $d$ , hay còn gọi là cập nhật các tham số cục bộ.
- Bước M: cập nhật lại tham số toàn cục  $\beta$ .

Bước E và bước M được lặp đi lặp lại cho đến khi các tham số hội tụ.

Trong [1], thay vì cập nhật trực tiếp ra các tham số cục bộ và toàn cục, tác giả cập nhật các tham số biến phân cho nó. Trong [6], tác giả Khoát Thân và Tùng Đoàn đề xuất thuật toán ML-OPE (cập nhật trực tiếp tham số toàn cục  $\beta$ ) và thuật toán Online-OPE (cập nhật tham số biến phân  $\lambda$  cho  $\beta$ ).

Trong [6], khi làm việc với mô hình LDA, các tác giả đưa ra bài toán suy diễn cho văn bản  $d$  là:

$$\theta^* = argmax_{\theta \in \Delta_K} \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k$$

Khi đó  $f(\theta) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj} + (\alpha - 1) \sum_{k=1}^K \log \theta_k$ .

Đặt  $g_1(\theta) = \sum_j d_j \log \sum_{k=1}^K \theta_k \beta_{kj}$  ;

$$g_2(\theta) = (\alpha - 1) \sum_{k=1}^K \log \theta_k$$

Như vậy ta có  $f(\theta) = g_1(\theta) + g_2(\theta)$ .

Với mô hình LDA, trong thực tế thường gặp tham số  $\alpha < 1$  nên khi đó thành phần  $g_1$  là hàm lõm và  $g_2$  là hàm lồi, nên hàm mục tiêu  $f(\theta)$  có dạng hàm DC. Do đó bài toán tìm cực đại của hàm  $f(\theta)$  là bài toán NP-khó, không có các thuật toán lặp xác định giải quyết hiệu quả bài toán tối ưu cho  $f(\theta)$ . Do đó ý tưởng của phương pháp giải xấp xỉ ngẫu nhiên [3] được đưa vào sử dụng để giải bài toán suy diễn hậu nghiệm.

Lấy ý tưởng từ thuật toán Online Frank-Wolfe [4], tác giả Khoát Thân và các cộng sự đề xuất thuật toán tối ưu ngẫu nhiên Online Maximum a Posterior Estimation (OPE) để giải quyết bài toán này [5,6].

**Thuật toán 1: OPE (Online Maximum a Posterior Estimation)**  
**Input:** document  $d$  and model  $\{\beta, \alpha\}$   
**Output:**  $\theta^*$  that maximize  $f(\theta) = g_1(\theta) + g_2(\theta)$

---

Initialize  $\theta_1$  arbitrarily in  $\overline{\Delta_K} = \{x \in \mathbb{R}^K : \sum_{k=1}^K x_k = 1, x \geq \epsilon > 0\}$   
**for**  $t = 1, 2, \dots, \infty$  **do**  
 Pick  $f_t$  uniformly from  $\{g_1(\theta); g_2(\theta)\}$   
 $F_t := \frac{2}{t} \sum_{h=1}^t f_h$   
 $e_t := argmax_{x \in \overline{\Delta_K}} \langle F_t, x \rangle$   
 $\theta_{t+1} := \theta_t + \frac{e_t - \theta_t}{t}$   
**end for**

Thuật toán OPE được sử dụng để giải quyết bài toán suy diễn véc tơ tỉ lệ chủ đề  $\theta_d$  cho từng văn bản  $d$ , sau đó thuật toán OPE được sử dụng vào trong thuật toán học ML-OPE áp dụng cho quá trình học với LDA [6]. OPE là thuật toán lặp ngẫu nhiên. Tại mỗi bước lặp  $t$ , thuật toán OPE chọn ngẫu nhiên một trong hai đại lượng  $g_1(\theta)$  hoặc  $g_2(\theta)$  với xác suất bằng nhau, sau đó tính trung bình các đại lượng đã chọn được từ các bước trước để tạo thành chuỗi hàm xấp xỉ  $F_t(\theta)$ . Ta có  $F_t(\theta) \rightarrow f(\theta)$  khi  $t \rightarrow \infty$ . Tại mỗi bước lặp  $t$ , OPE cập nhật  $\theta_{t+1}$  theo  $\theta_t$ . Khi  $t \rightarrow \infty$  thì  $\theta_t \rightarrow \theta^*$  trong đó  $\theta^*$  là một điểm dừng của  $f(\theta)$ .

Như vậy, thuật toán OPE sử dụng ý tưởng của thuật toán Online Frank-Wolfe: xây dựng dãy các hàm số  $F_t(\theta)$  để xấp xỉ một hàm mục tiêu  $f(\theta)$ . Cách xây dựng dãy hàm số này là: tại mỗi bước lặp, chọn ngẫu nhiên một trong hai thành phần  $\{g_1, g_2\}$ , hàm xấp xỉ bằng trung bình của tất cả các thành phần đã chọn tại các bước lặp trước.

**2. Một số ý tưởng cải tiến thuật toán OPE**

Chúng tôi nhận thấy một đặc điểm của hàm  $f(\theta)$  trong thuật toán OPE là  $g_1(\theta) < 0$ ,  $g_2(\theta) > 0$  với mọi giá trị của tham số  $\theta$ . Với cách xây dựng hàm  $F_t(\theta)$  như trên, nếu tại bước đầu tiên chọn được  $g_1(\theta)$  thì  $F_1 < f$ , nếu bước đầu tiên chọn được  $g_2(\theta)$  thì  $F_1 > f$ .

Hai dãy hàm có cùng cách xây dựng nhưng xuất phát khác nhau, cùng tiến đến hàm mục tiêu, một dãy bắt đầu từ  $g_1(\theta)$  (xuất phát từ phía dưới  $f(\theta)$  nên gọi là dãy  $L_t(\theta)$ ), một dãy bắt đầu từ  $g_2(\theta)$  (xuất phát từ phía trên  $f(\theta)$ , gọi là  $U_t(\theta)$ ). Hai dãy hàm số  $U_t(\theta)$  và  $L_t(\theta)$  sẽ có tương ứng hai dãy số  $\{\theta_t^u\}$  và  $\{\theta_t^l\}$  tiến dần đến điểm tối ưu  $\theta^*$  cần tìm. Chúng tôi tìm cách kết hợp hai dãy số này để được một dãy số tiến dần đến cực trị của hàm mục tiêu  $f(\theta)$ . Xây dựng hai dãy hàm ngẫu nhiên  $U_t(\theta), L_t(\theta)$  xuất phát từ bên trên và bên dưới hàm  $f$ , bao lấy hàm  $f$  và cùng tiến dần về  $f$ . Có nhiều sự lựa chọn trong quá trình xây dựng các biên ngẫu nhiên

$U_t(\theta), L_t(\theta)$  xấp xỉ cho hàm mục tiêu  $f(\theta)$  như sử dụng phân phối uniform, phân phối Bernoulli. Với việc xây dựng hai chuỗi hàm ngẫu nhiên  $U_t(\theta), L_t(\theta)$  bằng phân phối Bernoulli với tham số  $p \in (0,1)$  thích hợp tổng quát hơn phân phối đều, khi đó  $U_t(\theta)$  và  $L_t(\theta)$  là các xấp xỉ của  $f(\theta)$  khi  $t \rightarrow \infty$ . Cùng với việc xây dựng hai biên ngẫu nhiên, chúng tôi thu được hai dãy số  $\{\theta_t^u\}$  và  $\{\theta_t^l\}$  xấp xỉ cho dãy nghiệm  $\{\theta_t\}$ . Tại mỗi bước lặp chúng tôi lựa chọn  $\theta_t$  bằng cách sử dụng phân phối đều từ  $\theta_t^u$  và  $\theta_t^l$

$$\theta_t := \text{pick uniformly from } \{\theta_t^u; \theta_t^l\}$$

Chi tiết của ý tưởng này được trình bày trong Thuật toán 2, với tên là IOPE1.

<b>Thuật toán 2: IOPE1</b>	
<b>Input:</b>	document $d$ and model $\{\beta, \alpha\}$
<b>Output:</b>	$\theta^*$ that maximize $f(\theta) = g_1(\theta) + g_2(\theta)$
Initialize	$\theta_1$ arbitrarily in $\Delta_K = \{x \in \mathbb{R}^K : \sum_{k=1}^K x_k = 1, x \geq \epsilon > 0\}$
$f_1^u := g_1; f_1^l := g_2$	
<b>for</b> $t = 2, 3, \dots, \infty$ <b>do</b>	
Pick $f_t^u$ Bernoulli with probability $p$ from $\{g_1(\theta); g_2(\theta)\}$	
$U_t := \frac{1}{t} \sum_{h=1}^t f_h^u$	
$e_t^u := \text{argmax}_{x \in \Delta_K} \langle U_t^l(\theta_t), x \rangle$	
$\theta_{t+1}^u := \theta_t + \frac{e_t^u - \theta_t}{t}$	
Pick $f_t^l$ Bernoulli with probability $p$ from $\{g_1(\theta); g_2(\theta)\}$	
$L_t := \frac{1}{t} \sum_{h=1}^t f_h^l$	
$e_t^l := \text{argmax}_{x \in \Delta_K} \langle L_t^u(\theta_t), x \rangle$	
$\theta_{t+1}^l := \theta_t + \frac{e_t^l - \theta_t}{t}$	
$\theta_{t+1} := \text{pick uniformly from } \{\theta_{t+1}^u; \theta_{t+1}^l\}$	
<b>end for</b>	

Tiếp tục cải tiến thuật toán OPE theo hướng ngẫu nhiên hóa, chúng tôi nhận thấy sử dụng phân phối đều (uniform distribution) trong lựa chọn dãy nghiệm  $\{\theta_t\}$  thông qua  $\{\theta_t^u\}, \{\theta_t^l\}$  là đơn giản, chưa khai thác được giá trị của  $\{\theta_t^u\}, \{\theta_t^l\}$ . Vì vậy, chúng tôi sử dụng:

$$q := \frac{\exp(f(\theta_t^u))}{\exp(f(\theta_t^u)) + \exp(f(\theta_t^l))}$$

$\theta_t := \theta_t^u$  with probability  $q$

$\theta_t := \theta_t^l$  with probability  $1 - q$

Chi tiết được trình bày trong Thuật toán 3, với tên là IOPE2.

<b>Thuật toán 3: IOPE2</b>	
<b>Input:</b>	document $d$ and model $\{\beta, \alpha\}$
<b>Output:</b>	$\theta^*$ that maximize $f(\theta) = g_1(\theta) + g_2(\theta)$
Initialize	$\theta_1$ arbitrarily in $\Delta_K = \{x \in \mathbb{R}^K : \sum_{k=1}^K x_k = 1, x \geq \epsilon > 0\}$
	$f_1^u := g_1; f_1^l := g_2$
<b>for</b> $t = 2, 3, \dots \infty$ <b>do</b>	
	Pick $f_t^u$ Bernoulli with probability $p$ $\{g_1(\theta); g_2(\theta)\}$
	$U_t := \frac{1}{t} \sum_{h=1}^t f_h^u$
	$e_t^u := \operatorname{argmax}_{x \in \Delta_K} \langle U_t^l(\theta_t), x \rangle$
	$\theta_{t+1}^u := \theta_t + \frac{e_t^u - \theta_t}{t}$
	Pick $f_t^l$ Bernoulli with probability $p$ $\{g_1(\theta); g_2(\theta)\}$
	$L_t := \frac{1}{t} \sum_{h=1}^t f_h^l$
	$e_t^l := \operatorname{argmax}_{x \in \Delta_K} \langle L_t^l(\theta_t), x \rangle$
	$\theta_{t+1}^l := \theta_t + \frac{e_t^l - \theta_t}{t}$
	$q := \frac{\exp(f(\theta_{t+1}^u))}{\exp(f(\theta_{t+1}^u)) + \exp(f(\theta_{t+1}^l))}$
	$\theta_{t+1} := \theta_{t+1}^u$ with probability $q$
	$\theta_{t+1} := \theta_{t+1}^l$ with probability $1 - q$
<b>end for</b>	

Với việc đưa ra hai cải tiến thuật toán IOPE1 và IOPE2 cho bài toán suy diễn hậu nghiệm với LDA, chúng tôi áp dụng vào thuật toán học với LDA và đưa ra thuật toán học ML-IOPE1 và ML-IOPE2 trong thuật toán 4.

<b>Thuật toán 4: ML-IOPE1 (2)</b>	
<b>Input:</b>	data sequence, $K, \alpha, \tau > 0, \kappa \in (0.5, 1]$
<b>Output:</b>	$\beta$
Initialize	$\beta^0$ randomly in $\Delta_V$
<b>for</b> $t = 1, 2, \dots \infty$ <b>do</b>	
	Pick a set $C_t$ of documents
	Do inference by IOPE1(2) for each $d \in C_t$ to get $\theta_d$ , given $\beta^{t-1}$
	Compute intermediate topics $\hat{\beta}^t$ as:
	$\hat{\beta}_{kj}^t \propto \sum_{d \in C_t} d_j \theta_{dk}$
	Set step-size: $\rho_t := (t + \tau)^{-\kappa}$
	Update topics: $\beta^t := (1 - \rho_t)\beta^{t-1} + \rho_t \hat{\beta}^t$
<b>end for</b>	

### 3. Thử nghiệm

#### 3.1 Các độ đo

3.1.1 Độ đo khả năng tổng quát hóa của mô hình Log Predictive Probability: Xác suất dự đoán (Predictive Probability) chỉ ra khả năng tổng quát hóa của mô hình  $\mathcal{M}$  trên một dữ liệu mới. Dữ liệu mới ở đây là một văn bản  $w$ . Với một văn bản mới, ta chia văn bản thành 2 phần  $w_{obs}$  và  $w_{ho}$  với tỉ lệ  $w_{obs}:w_{ho} = 80:20$ . Tiếp theo ta suy diễn cho phần  $w_{obs}$  để ước lượng  $E(\theta^{obs})$ . Sau đó ước lượng Predictive Probability:

$$\begin{aligned} \Pr(w_{ho} | w_{obs}, \mathcal{M}) &\approx \prod_{w \in w_{ho}} \sum_{k=1}^K E(\theta_k^{obs}) E(\beta_{kw}) \\ \text{Log Predictive Probability} &= \frac{\log \Pr(w_{ho} | w_{obs}, \mathcal{M})}{|w_{ho}|} \end{aligned}$$

Với  $\mathcal{M}$  là mô hình được tính độ đo perplexity.

3.1.2 Độ đo chất lượng chủ đề NPMI: Độ đo NPMI [2] nói về chất lượng của từng chủ đề học được. Với các mô hình chủ đề, độ đo NPMI đánh giá tương đối tốt với suy diễn của con người trong một chủ đề. Với mỗi chủ đề  $t$ , ta chọn ra  $n$  từ có xác suất cao nhất  $\{w_1, w_2 \dots w_n\}$  và tính độ đo NPMI của chủ đề đó:

$$NPMI(t) = \frac{2}{n(n-1)} \sum_{j=2}^n \sum_{i=1}^{j-1} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}$$

Trong đó  $P(w_i, w_j)$  là xác suất để từ  $w_i$  và  $w_j$  cùng xuất hiện trong một văn bản.

Với mô hình  $\mathcal{M}$  có  $K$  chủ đề, độ đo NPMI trên mô hình này được tính như sau:

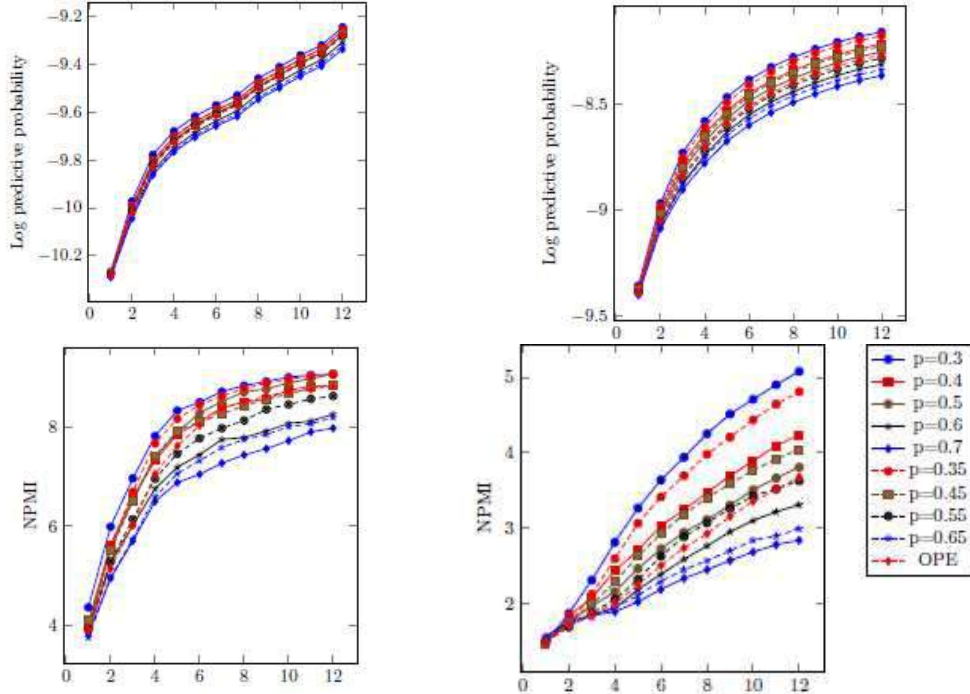
$$NPMI = \frac{1}{K} \sum_{t=1}^K NPMI(t)$$

#### 3.2 Mô tả thử nghiệm

Trình bày về các kết quả thử nghiệm trên các bộ dữ liệu thực tế của các cải tiến và so sánh với thuật toán OPE.

**Bảng 1. Dữ liệu thử nghiệm**

Bộ dữ liệu	Số lượng văn bản	Tập từ điển
New York Times	300.000	102661
Pubmed	310.000	141044



a. Với bộ dữ liệu New York Times

b. Với bộ dữ liệu PubMed

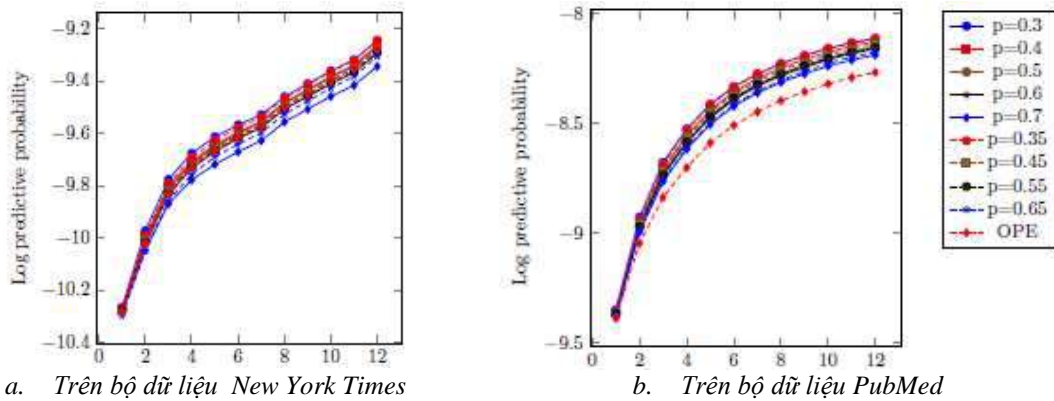
**Hình 2. Kết quả của phương pháp học ML-IOPE1 với các giá trị Bernoulli p khác nhau**

Thuật toán IOPE1, IOPE2 được sử dụng để áp dụng vào thuật toán ML-OPE học LDA. Ta đi so sánh các cải tiến khi được thay thế cho OPE trong thuật toán trên. OPE và các cải tiến là các thuật toán ngẫu nhiên, do đó ta sẽ chạy các thuật toán 5 lần, lấy trung bình kết quả các lần chạy và so sánh với nhau.

Tham số của mô hình:

$$K = 100, \alpha = \frac{1}{K}, \eta = \frac{1}{K}, \text{ số lần lặp } T = 50, \kappa = 0,9, \tau = 1.$$

Thông qua kết quả thử nghiệm, ta thấy với tham số p của phân phối Bernoulli được lựa chọn thích hợp ta thấy thuật toán IOPE1, IOPE2 hiệu quả hơn cả thuật toán OPE (mặc dù OPE đã được đánh giá tốt hơn các thuật toán hiện có như VB, CVB hay CGS [5,6]). Đặc biệt khi lựa chọn tham số Bernoulli  $p < 0.5$  như trong thực nghiệm  $p = 0.3, 0.35, 0.4$  thì kết quả tốt hơn OPE rất nhiều trên cả hai độ đo Log Predictive Probability và NPMI với hai bộ dữ liệu New York Times và Pubmed.



Hình 3. Kết quả của phương pháp học ML-IOPE2 với các giá trị Bernoulli  $p$  khác nhau

#### 4. Kết luận

Sử dụng cách tiếp cận ngẫu nhiên, bài báo đưa ra các thuật toán tối ưu hiệu quả giải bài toán suy diễn hậu nghiệm dưới dạng một tối ưu không lồi khó giải. Bài báo đưa ra hai thuật toán cải tiến là IOPE1 và IOPE2 bằng việc sử dụng phân phối Bernoulli cải tiến cho thuật toán OPE với tham số Bernoulli  $p$  thích hợp. Kết quả thử nghiệm trên hai bộ dữ liệu lớn là New York Times và Pubmed cho thấy cho thấy các thuật toán đề xuất là hiệu quả so với các kết quả đã có, đặc biệt khi chúng ta lựa chọn tham số Bernoulli  $p$  phù hợp. Tham số  $p$  giúp thuật toán thể hiện tính linh hoạt và đóng vai trò là tham số hiệu chỉnh của thuật toán.

#### TÀI LIỆU THAM KHẢO

- [1]. David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, 3(3), pp. 993–1022, 2003.
- [2]. Nikolaos Aletras and Mark Stevenson, "Evaluating topic coherence using distributional semantics", *In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pp. 13-22, 2013.
- [3]. Léon Bottou, "Online learning and stochastic approximations", *Online learning in neural networks*, 17(9), pp.142, 1998
- [4]. Elad Hazan, Satyen Kale, *Projection-free Online Learning*, ICML 2012.
- [5]. Khoat Than, Tung Doan, "Dual online inference for latent Dirichlet allocation", *In ACML. Journal of Machine Learning Research: W&CP*, 2014.
- [6]. Khoat Than, Tung Doan, *Fast algorithms for inference in topic models*, Technical report, 2016.