Vol.2009-NL-193 No.7
2009/9/28

Latent Dirichlet Allocation における決定論的オンラインベイズ学習

佐藤一誠[†1]　　中川裕志[†2]

近年、機械学習やデータマイニングにおいて、トピックモデルと呼ばれる確率的生成モデルの研究が盛んに行われている．本研究では，代表的なトピックモデルの 1 つである Latent Dirichlet allocation (LDA) のオンライン学習手法を提案する．従来の LDA のオンライン学習手法は，サンプリング手法に基づいており，大量のテキストストリームデータに対して，学習速度や収束性の保証の問題がある．また，トピック分布の分布である Dirichlet 分布のパラメータ推定をオンラインで学習していないという問題があった．本研究では，Dirichlet 分布のパラメータを含め，LDA におけるパラメータのオンライン学習を行い，その局所解への収束性を保証する．また，文書モデルの実験において一括 (batch) 学習と同程度の性能があることを示す．

Online learning for Latent Dirichlet Allocaiton

ISSEI SATO[†2] and HIROSHI NAKAGAWA[†2]

One of the important approaches for Knowledge discovery and Data mining is to estimate unobserved variables because latent variables can indicate hidden and specific properties of observed data. Latent Dirichlet allocation(LDA) plays a important role in unobserved document modeling where latent variables indicate topics in documents. We introduce online learning algorithm for LDA based on variational Bayes in the case that documents arrive in a continuous stream and a large number of documents are accumulated. In an experiment using real data, this online method performs as well as batch learning in LDA.

†1 東京大学大学院 情報理工学系研究科
Graduate School of Information Sience and Technology, The University of Tokyo
†2 東京大学 情報基盤センター
Interfaculty Initiative in Information Studies

1. はじめに

近年、機械学習やデータマイニングにおいて、トピックモデルと呼ばれる確率的生成モデルの研究が盛んに行われている．トピックとは，データの隠れた情報や性質を表わす潜在的なクラスである．例えば，文書データの場合，トピックとは文書に潜む分野情報もしくは単語の意味カテゴリとなる．Latent Dirichlet allocation (LDA)[1] は，代表的なトピックモデルの 1 つである．LDA は，トピックの分布を多項 (multinomial) 分布でモデル化し，トピック分布の分布に対して Dirichlet 分布を仮定する．Dirichlet 分布は，多項分布の共役事前分布であるため，学習アルゴリズムの導出の容易さからも自然なモデル化となっている．LDA をベースとするトピックモデルは，さまざまな形で拡張されており，例えば，文法的な要素を考慮したモデル[2)3)]，時系列データのモデル化[4),5)]，著者や citation 情報の利用[6)–8)]，Hypertext のモデル化[9)]，視覚化への応用[10)] などがある．

トピックモデルの学習は多くの場合，一括 (batch) 学習である．一括学習では学習用データが一括ですべて与えられ，パラメータの学習を行う．一方，オンライン学習 (逐次学習) では，データは 1 つずつ逐次的に与えられ，データが与えられるたびにパラメータを更新する．一般的にオンライン学習の長所としては，

- 全てのデータを蓄積する必要がなく，少メモリで大規模なデータを扱える
- 増加したデータに対してのみ学習し直せば良いので，すべてのデータから学習する一括学習より計算量が少なくて済む

ということが，挙げられる．しかしながら，

- 学習速度と正しい解への収束性のトレードオフがある
- オンライン学習の手法によっては，パラメータの更新の際に用いる学習率と呼ばれるパラメータの設定に性能が依存する

という問題点もある．

トピックモデルは主に教師なし学習であるため，扱うデータ量は膨大で，日々増加するデータに対して一括学習を行うのは現実的ではない．したがって，オンライン学習が適していると考えられる．Canini ら[11)] は，Incremental Gibbs sampler や Particle filter などのサンプリング手法に基づく LDA の学習手法を提案している．しかし，サンプリングに基づくオンライン学習では，大量のテキストストリームデータに対して，学習速度や収束性の保証がないという問題がある．また，トピック分布の分布である Dirichlet 分布のパラメータ推定をオンラインで学習していないという問題がある．これに対して，本稿で提案する LDA のオ

ンライン学習手法は以下の特徴をもつ.

- 変分ベイズ法に基づく決定論的学習手法である.
- トピック分布上の分布である Dirichlet 分布のパラメータのオンライン学習を行う.
- パラメータの学習に対して局所解への収束を保証する.
- 学習率は，ベイズ学習の枠組み内で導出される．具体的には，学習率の設定が，ベイズ学習における事前分布のパラメータの決定と等価である.

以下, 第 2 節では LDA について説明し, 第 3 節では, 提案する LDA でのオンライン学習手法について説明する．第 4 節で実験結果について説明する.

2. LDA

本節では, LDA について簡単に説明する. LDA は, 単語頻度分布によって特徴づけられる潜在トピックと呼ばれる隠れ変数を用いて文書の生成過程をモデル化する.

2.1 LDA による文書の生成過程

本論文における記号を以下に示す. T をトピック数とする. M を文書数とする. V は, 語彙数を示す. N_j は文書 j における全単語数である. $w_{j,i}$ は, 文書 j における i 番目の単語を示す. $z_{j,i}$ は, 単語 $w_{j,i}$ における潜在トピックである. $Multi(\cdot)$ を多項 (multinomial) 分布とする. $Dir(\cdot)$ を Dirichlet 分布とする. $\boldsymbol{\theta}_j$ は, T 次元確率ベクトルで, 文書 j におけるトピック分布を示す. $\boldsymbol{\beta}$ は $T \times V$ 行列で, $\beta_{t,v}$ は, トピック t 単語 v の出現確率を示す. $\boldsymbol{\alpha}$ は, T 次元の Dirichlet 分布のパラメータである.

LDA の生成モデルを以下示す.

- For each of the M documents j:
  - Draw $\boldsymbol{\theta}_j \sim Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})$,
  - For each of the N_j words $w_{j,i}$:
    - Draw topic $z_{j,i} \sim Multi(z|\boldsymbol{\theta}_j)$,
    - Draw word $w_{j,i} \sim p(w|z_{j,i}, \boldsymbol{\beta})$,

ここで, $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_t \theta_t^{\alpha_t - 1}$, $p(w = v|z = t, \boldsymbol{\beta}) = \beta_{t,v}$.

2.2 LDA の変分ベイズ法による学習

LDA の変分ベイズ法[1] は，以下のように近似事後分布 $q(\boldsymbol{z}, \boldsymbol{\theta})$ を導入する.

$$q(\boldsymbol{z}, \boldsymbol{\theta}) = \prod_{j,i} q(z_{j,i}|\boldsymbol{\phi}_{j,i}) \prod_j q(\boldsymbol{\theta}_j|\boldsymbol{\gamma}_j), \tag{1}$$

ここで,$\boldsymbol{\phi}$,$\boldsymbol{\gamma}$ は変分パラメータである．$\phi_{j,i,t}$ は, 文書 j における i 番目の単語 $w_{j,i}$ のトピックが t であることを示す. $\boldsymbol{\gamma}_j$ は $\boldsymbol{\theta}_j$ 上の Dirichlet 分布のパラメータである, すなわち, $q(\boldsymbol{\theta}_j|\boldsymbol{\gamma}_j) \propto \prod_t \theta_{j,t}^{\gamma_{j,t}-1}$. 変分ベイズ法では, 事後分布 $p(\boldsymbol{z}, \boldsymbol{\theta}|cD)$ と近似事後分布 $q(\boldsymbol{z}, \boldsymbol{\theta})$ との間の Kullback-Leibler 情報量 $KL[q(\boldsymbol{z}, \boldsymbol{\theta})|p(\boldsymbol{z}, \boldsymbol{\theta}|cD)]$ が最少になるような各パラメータの更新式を求める.

実際には，文書集合 $\mathcal{D}$ における以下の対数尤度に対して，近似事後分布を導入することで得られる下限を最大化することで，パラメータの更新式を求める.

$$\mathcal{LL}_{LDA}(\mathcal{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_j^M \log \int \sum_{\boldsymbol{z}} \prod_i^{N_j} p(w_{j,i}|z_{j,i}, \boldsymbol{\beta}) p(\boldsymbol{z}_j|\boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j|\boldsymbol{\alpha}) d\boldsymbol{\theta}_j \tag{2}$$

$$\phi_{j,i,t} \propto \beta_{t w_{j,i}} \exp\left\{\Psi(\gamma_{j,t} - \Psi(\sum_{t=i}^{T} \gamma_{j,t}))\right\}, \tag{3}$$

$$\gamma_{j,t} = \alpha_t + n_{j,t}, \quad n_{j,t} = \sum_{i=1}^{N_j} \phi_{j,i,t} \tag{4}$$

$$\beta_{t,v} \propto \lambda_0 + \sum_j^M n_{j,t,v}, \quad n_{j,t,v} = \sum_i^{N_j} \phi_{j,i,t} \mathbb{I}(w_{j,i} = v). \tag{5}$$

$\mathbb{I}(\cdot)$ を, 指示関数 (indicator function) とする. λ_0 は,$\{\boldsymbol{\beta}_t\}$ 上の Dirichle 分布のパラメータで, トピックごとの単語分布におけるスムージング項となっている. $\boldsymbol{\alpha}$ は, Newton-Raphson 法[1] や固定点反復法[12] によって推定する. LDA の学習アルゴリズムを Algorithm 1 に示す.

2.3 Dirichlet 分布のパラメータ $\boldsymbol{\alpha}$ の推定

Blei らは,Dirichlet 分布のパラメータ $\boldsymbol{\alpha}$ を, Newton-Raphson 法を用いて推定した[1]. しかしながら，Newton-Raphson 法は収束が遅く，初期値によっては負の値をとることもあることから，固定点反復法[12] を用いるほうが効率的であることが知られている. 固定点反復法では，固定点方程式を導出する必要がある．LDA では，次の評価式[12] を用いることにより,

$$\frac{\Gamma(x)}{\Gamma(n+x)} \geq \frac{\Gamma(\hat{x}) \exp((\hat{x} - x)b)}{\Gamma(n + \hat{x})}, \frac{\Gamma(n+x)}{\Gamma(x)} \geq cx^a \quad n > 0, \tag{6}$$

$$a = \{\Psi(n + \hat{x}) - \Psi(\hat{x})\}x, b = \Psi(n + \hat{x}) - \Psi(\hat{x}), c = \frac{\Gamma(n + \hat{x})}{\Gamma(\hat{x})} \hat{x}^{-a}. \tag{7}$$

以下の固定点方程式を導出することができる.

$$\alpha_t = \frac{\sum_j \{\Psi(\alpha_t + n_{j,t}) - \Psi(\alpha_t)\}\alpha_t}{\sum_j (\Psi(N_j + \alpha_0) - \Psi(\alpha_0))}, \quad \alpha_0 = \sum_t \alpha_t. \tag{8}$$

Vol.2009-NL-193 No.7
2009/9/28

Algorithm 1 LDA における変分ベイズ法

Set L, which is the number of iterations.
Initialize $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and ϕ.
for all iterations it such that $1 \leq it \leq L$ **do**
  for $j = 1, ..., M$ **do**
    for $i = 1, ..., N_j$ **do**
      Update $\phi_{j,i,t}$ $(t = 1, \cdots, T)$ by Eq. (3)
    end for
    Update $\gamma_{j,t}$ $(t = 1, \cdots, T)$ by Eq. (4)
  end for
  Update β by Eq. (5)
  Update $\boldsymbol{\alpha}$ by the Newton-Raphson method[1] or fixted point iteration[12].
end for

3. LDA における決定論的オンライン学習

3.1 教師なしオンライン学習

LDA の学習は一般に教師なし学習である．教師なし学習におけるオンライン学習は主に Incremental 学習[13] と Stepwise 学習[14] がある．これらはともに EM アルゴリズムに対して提案されているが変分ベイズでも同様に扱うことができる．Liang ら[15] は，この 2 つのオンラインアルゴリズムと一括学習との比較をいくつかの自然言語処理のタスクにおける教師なし学習に対して行っている．教師なし学習では，パラメータ σ の更新を各データ $(i = 1, ..., M)$ から推定される統計量 s_i を用いて行われる．

一括学習において次のように各データの統計量の和によってパラメータ更新をここなう場合，オンライン化することは容易である．

(1) Initialize σ
(2) for each iteration
  (a) for each data $i = 1, ..., M$,
    (i) inference s_i by σ //EM アルゴリズムの E ステップに相当する
  (b) update $\sigma = \sum_i s_i$ //EM アルゴリズムの M ステップに相当する

Algorithm 2 LDA におけるオンライン学習

Set l, which is the number of inner iterations.
Initialize $\boldsymbol{\alpha}$
for $j = 1, ..., M$ **do**
  Initialize $\boldsymbol{\phi}_j$ and compute $n_{j,t}$, $n_{j,t,v}$, $\gamma_{j,t}$ and β.
  for all iterations it such that $1 \leq it \leq l$ **do**
    for $i = 1, ..., N_j$ **do**
      Update $\phi_{j,i,t}$ $(t = 1, \cdots, T)$ by Eq. (9)
      Update $n_{j,t}$ $(t = 1, \cdots, T)$ by Eq. (10)
      Update $n_{j,t,w_{j,i}}$ by Eq. (11)
      Update $\gamma_{j,t}$ $(t = 1, \cdots, T)$ by Eq. (12)
      Update β by Eq. (13)
    end for
  end for
  Update $\boldsymbol{\alpha}$ by Eq.(18)
end for

Incremental 学習
(1) Initialize s_i $(i = 1, ..., M)$.
(2) $\sigma = \sum_i s_i$.
(3) for each iteration
  (a) for each data $i = 1, ..., M$,
    (i) inference s_i by σ,
    (ii) $\sigma^{new} = \sigma^{old} - s_i^{old} + s_i^{new}$.

Stepwise 学習 (η_i は学習率)
(1) Initialize σ
(2) for each iteration
  (a) for each data $i = 1, ..., M$,
    (i) inference s_i by σ,
    (ii) $\sigma^{new} = (1 - \eta_i)\sigma^{old} + \eta_i s_i^{new}$.

Incremental 学習の特徴は，

- 学習率を設定する必要がない，
- すべてのデータに対して過去の統計量 s_i^{old} を保持しなければならない．

その反対に，Stepwise 学習は，

- 統計量 s_i^{old} を保持する必要がない，
- 学習率を設定する必要がある，

- 学習率の設定によって時系列的なトピックドリフトを扱うことができる．

という特徴がある．

オンライン学習は, 各データごとにパラメータ更新を行うことから，すべてのデータに対して 1 度だけパラメータ更新しを行う一括学習に比べ，収束が速い場合が多い．

3.2 問題設定と学習アルゴリズムの導出

本研究で扱うデータは，日々増加するテキストストリームデータである．したがって，前節で説明した Incremental 学習や Stepwise 学習は，全データに対して反復 (Iteration) を仮定しているが，本研究では，反復を仮定しない．つまり，全文書集合に対しての反復を行わず，各文書が与えられ，学習が終わったら次々に捨てていくものとする．もし，一度学習したデータを再び学習する必要がある場合は，新たなデータとして扱うことにする．ただし，本研究で扱うデータは文書データで，1 文書を 1 つの単位とすることから，1 文書内での各単語のトピックの推定では反復を行う．

具体的なアルゴリズムを Algorithm 2 に示す文書内での反復 (Algorithm 2 の 5-13) では，Incremental 学習を行っていることになる．ただし，1 文書内での反復なので，古い統計量を保持しておく必要は 1 文書内の単語に対応する統計量のみである．次の文書を学習する場合，これらの統計量は捨ててしまうため Stepwise 学習と同様にデータすべてに対して古い統計量を保持しておく必要はない．また，Incremental 学習は，対数尤度の下限を単調に増加することが保証されているため[13]，文書単位での対数尤度の下限の単調増加が保証できる．

一括学習と同様に，与えられたパラメータのもとで，各文書における各単語ごとにトピックへの帰属確率を計算する．

$$\phi^{new}_{j,i,t} \propto \beta_{tw_{j,i}} \exp\left\{\Psi(\gamma_{j,t} - \Psi(\sum_{t=i}^{T} \gamma_{j,t}))\right\} \tag{9}$$

各単語ごとに以下の統計量を Incremental 学習で更新する．

$$n^{new}_{j,t} = n^{old}_{j,t} - \phi^{old}_{j,i,t} + \phi^{new}_{j,i,t}, \tag{10}$$

$$n^{new}_{j,t,v} = n^{old}_{j,t,v} + (-\phi^{old}_{j,i,t} + \phi^{new}_{j,i,t})\mathbb{I}(w_{j,i} = v) \tag{11}$$

上記の統計量をもとにパラメータを単語ごとに更新する．

$$\gamma^{new}_{j,t} = \alpha_t + n^{new}_{j,t}, \tag{12}$$

$$\beta^{new}_{t,v} \propto \lambda_0 + \sum_{d}^{j-1} n_{d,t,v} + n^{new}_{j,t,v}. \tag{13}$$

3.3 パラメータ $\boldsymbol{\alpha}$ のオンライン推定

3.1 節で示した通り，各データの統計量 $s_i (i = 1, ..., M)$ の和によってパラメータ更新を行う場合，オンライン化することは容易であった. 実際，前節において，パラメータ $\gamma_{j,t}$，$\beta_{t,v}$ は，各文書から得られる統計量 $n_{j,t}$，$n_{j,t,v}$ もしくは $\phi_{j,i,t}$ の和になっているため容易にオンラインでの更新式を導出できた. しかし，パラメータ $\boldsymbol{\alpha}$ の更新式は，Eq.(8) からわかる通り，単純な統計量の和になっていない．本節では，まず，Eq.(8) による α_t の更新式が，α_t にガンマ分布を仮定した場合の期待値計算になっていることを示す．次に，ガンマ分布のパラメータをオンラインで更新することによって，α_t をオンラインで推定できることを示す．

α_t に対してガンマ分布 $G(\alpha|a_0, b_0)$ を仮定する，すなわち，$\alpha_t \sim G(\alpha|a_0, b_0)(t = 1, ..., T)$. ここで，$a_0$，$b_0$ はガンマ分布のパラメータである．

事前分布としてガンマ分布を導入することにより，Eq.(8) で与えられる更新式は，Eq.(7) を用いて以下のようになる．

$$\alpha^{new}_t = \frac{a_0 - 1 + \sum_j^M a_{j,t}}{b_0 + \sum_j^M b_j}, \tag{14}$$

$$a_{j,t} = \{\Psi(\alpha^{old}_t + n_{j,t}) - \Psi(\alpha^{old}_t)\}\alpha^{old}_t, \quad b_j = \Psi(N_j + \alpha^{old}_0) - \Psi(\alpha^{old}_0) \tag{15}$$

ここで，データが与えられた下でのガンマ分布の事後分布を $G(\alpha|\tilde{a}_t, \tilde{b})$ とすれば

$$\mathbb{E}[\alpha_t]_{G(\alpha|\tilde{a}_t,\tilde{b})} = \frac{\tilde{a}_t}{\tilde{b}}, \tag{16}$$

$$\tilde{a}_t = a_0 - 1 + \sum_j^M a_{j,t}, \quad \tilde{b} = b_0 + \sum_j^M b_j \tag{17}$$

となり，α_t の更新式 (14) になっていることがわかる．したがって，$a_{j,t}$,b_j は各々のデータ j から得られる統計量とみなすことができ，われわれがオンラインで更新するパラメータは $\tilde{a}_t$，$\tilde{b}$ であることがわかる．これらは，各データから得られる統計量の和によって求めるため容易にオンラインの更新式を導出することができる．j 番目のデータが得られた下でのパラメータ $\alpha^{(j)}_t$ のオンライン推定は以下となる．

$$\alpha^{(j)}_t = \mathbb{E}[\alpha_t]_{G(\alpha|\tilde{a}^{(j)}_t,\tilde{b}^{(j)})} = \frac{\tilde{a}^{(j)}_t}{\tilde{b}^{(j)}}, \tag{18}$$

$$\tilde{a}^{(j)}_t = a_0 - 1 + \sum_d^{j-1} a_{d,t} + a_{j,t}, \quad \tilde{b}^{(j)} = b_0 + \sum_d^{j-1} b_d + b_j \tag{19}$$

$$a_{j,t} = \{\Psi(\alpha^{(j-1)}_t + n_{j,t}) - \Psi(\alpha^{(j-1)}_t)\}\alpha^{(j-1)}_t, \quad b_j = \Psi(N_j + \alpha^{(j-1)}_0) - \Psi(\alpha^{(j-1)}_0) \tag{20}$$

3.4 パラメータ学習の収束性

本研究で提案する LDA のオンライン学習では，文書単位で学習の反復を行う場合 Incremental 学習を行っていた．本節では，j 番目のデータが得られた下でのパラメータ $\boldsymbol{\alpha}^{(j)}$，$\boldsymbol{\beta}^{(j)}$

のオンラインの更新式が，Stepwise 学習におけるパラメータ更新式となっていることを示す. 次に，Stepwise 学習の枠組みの中で $\boldsymbol{\alpha}^{(j)}$，$\boldsymbol{\beta}^{(j)}$ のオンライン学習が局所最適解へ収束することを示す. 変分ベイズに基づくトピックモデルの決定論的学習では，正しい解を見つけることは難しく，局所最適解への収束性が保証されれば十分である.

Eq.(18) より

$$\alpha_t^{(j)} = \frac{a_0 - 1 + \sum_d^{j-1} a_{d,t}}{b_0 + \sum_d^{j-1} b_d}(1-\eta_j^\alpha) + \eta_j^\alpha \frac{a_{j,t}}{b_j} = \alpha_t^{(j-1)}(1-\eta_j^\alpha) + \eta_j^\alpha \frac{a_{j,t}}{b_j} \tag{21}$$

$$\eta_j^\alpha = \frac{b_j}{b_0 + \sum_d^j b_d} \tag{22}$$

Eq.(13) より

$$\beta_{t,v}^{(j)} = \frac{\lambda_0 + \sum_d^{j-1} n_{d,t,v}}{V\lambda_0 + \sum_d^{j-1} n_{d,t,\cdot}}(1-\eta_j^\beta) + \eta_j^\beta \frac{n_{j,t,v}}{n_{j,t,\cdot}} = \beta_{t,v}^{(j-1)}(1-\eta_j^\beta) + \eta_j^\beta \frac{n_{j,t,v}}{n_{j,t,\cdot}} \tag{23}$$

$$\eta_j^\beta = \frac{n_{j,t,\cdot}}{V\lambda_0 + \sum_d^{j-1} n_{d,t,\cdot} + n_{j,t,\cdot}} \tag{24}$$

3.1 節で説明した通り，この α，β の更新式は η_j^α,η_j^β を学習率とする Stepwise 学習になっている. 通常，Stepwise 学習においては，学習率を人手で与えるが，本研究では，上記のように導出したものを用いる. 上記の導出において，学習率は，ハイパーパラメータ b_0，λ_0 によって定まることから，ベイズ学習において事前分布をどのように設定するかという問題と等価である.

次に，パラメータ学習の収束性について説明する．学習率 η_j^α,η_j^β は，以下のように抽象化して書くことができる.

$$\eta_j = \frac{S_j}{\tau + \sum_d^j S_d} \tag{25}$$

したがって，パラメータ学習の収束性を Eq.(25) の η_j について議論する.

η_j が以下を満たすとき, 局所解に収束することが保証されている[16)]

$$\sum_j^\infty \eta_j = \infty, \quad \sum_j^\infty \eta_j^2 < \infty \tag{26}$$

また, 以下の学習率の場合, 条件 (26) を満たすことが知られている.

$$\eta_j = \frac{\tau_1}{\tau_2 + j} \ (\tau_1, \tau_2 > 0) \tag{27}$$

したがって，以下の定理 1 によりパラメータ $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ が提案するオンライン学習において局所解へ収束性することが保証される.

Theorem 1. 任意の j に対して,$\epsilon < S_j < \nu$ を満たす ϵ,ν (> 0) が存在するとき, 以下の学習率は, 条件 (26) を満たす.

$$\eta_j = \frac{S_j}{\tau + \sum_d^j S_d} \tag{28}$$

Proof. $0 < \epsilon < S_j < \nu$ を満たす ϵ,ν が存在するとき

$$\frac{\epsilon}{\tau + j\nu} < \eta_j = \frac{S_j}{\tau + \sum_d^j S_d} < \frac{\nu}{\tau + j\epsilon} \tag{29}$$

Eq.(27) は条件 (26) を満たすことから

$$\sum_j^\infty \frac{\epsilon}{\tau + j\nu} = \sum_j^\infty \frac{\epsilon/\nu}{\tau/\nu + j} = \infty, \quad \sum_j^\infty \frac{\nu}{\tau + j\epsilon} = \sum_j^\infty \frac{\nu/\epsilon}{\tau/\epsilon + j} = \infty \tag{30}$$

はさみうちの原理から

$$\sum_j^\infty \frac{S_j}{\tau + \sum_d^j S_d} = \infty \tag{31}$$

また

$$\sum_j^\infty \left(\frac{S_j}{\tau + \sum_d^j S_d}\right)^2 < \sum_j^\infty \left(\frac{\nu/\epsilon}{\tau/\epsilon + j}\right)^2 < \infty \tag{32}$$

したがって、Eq.(28) の η_j は条件 (26) を満たす. □

4. 実　　験

オンライン学習の性能を以下の 2 つのコーパスを用いて評価する: TREC[⋆1] で用いられている “AP(Associated Press)” コーパス ($M = 10,000, V = 67,291$) と “WSJ(The Wall Street Journal)” コーパス ($M = 10,000, V = 56,738$). 頻度 1 の語及びストップワードを除去した. 各々訓練文書とは別に，テスト文書 (1,000 文書) の Perplexity(低いほど良い) で評価した．訓練文書の順番は時系列に学習した．初期値を変えた 5 回の実験の平均と 2 乗誤差をトピックごとに計算したが，誤差をグラフに反映させたところほとんど見分けがつかなかったため，グラフには平均のみ載せている．事前分布のパラメータをそれぞれ $\lambda_0 = 100/V$, $a_0 = 1, b_0 = 1$ とした．一括学習を LDA として，Algorithm1 における反復回数 L を変えて実験を行った．提案するオンライン学習を OLDA とし，Algorithm2 における文書内での反復回数 l を変えて実験を行った．図 1 に実験結果を示す．提案するオンライン学習は一括学習と同程度の性能があることがわかる．さらに，一括学習は，トピック数が多い場合，反復回数が不十分となり性能が劣化する現象がみられるが，オンライン学習では劣化が起こ

⋆1 http://trec.nist.gov/

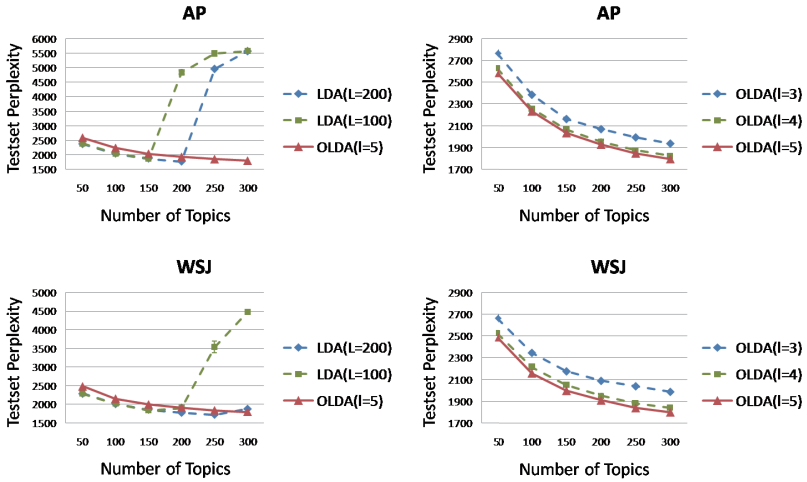

図 1 実験結果

らない．また，$l = 5$ 程度で収束していることもわかる．学習時間については，トピック数 $T = 300$ の場合，LDA($L = 100$) では，AP で，61,810[s], WSJ で，50,118[s] であるのに対し，OLDA($l = 5$) では，AP で 3,690[s]，WSJ で，3,075[s] であった．

5. おわりに

本研究では，LDA における決定論的オンラインベイズ学習を提案した．提案したアルゴリズムは，パラメータの学習においての局所解への収束性を保証する．また，文書モデルの実験において一括学習と同程度の性能があることを示した．

謝　辞

本研究は，文科省科学研究費 特定領域研究「情報爆発」の補助を得て行われた．

参 考 文 献

1) D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
2) T.L. Griffiths, M.Steyvers, D.M. Blei, and J.B. Tenenbaum. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, 2005.
3) J.L. Boyd-Graber and D.Blei. Syntactic Topic Models. In *NIPS*, 2008.
4) D.M. Blei and J.D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, New York, NY, USA, 2006. ACM.
5) C.Wang, D.M. Blei, and D.Heckerman. Continuous Time Dynamic Topic Models. In *UAI*, pages 579–586, 2008.
6) M.Rosen-Zvi, T.Griffiths, M.Steyvers, and P.Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, Arlington, VA, USA, 2004. AUAI Press.
7) M.Steyvers, P.Smyth, M.Rosen-Zvi, and T.Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, New York, NY, USA, 2004. ACM Press.
8) R.M. Nallapati, A.Ahmed, E.P. Xing, and W.W. Cohen. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550, New York, NY, USA, 2008. ACM.
9) A.Gruber, M.Rosen-Zvi, and Y.Weiss. Latent Topic Models for Hypertext. In *UAI*, pages 230–239, 2008.
10) T.Iwata, T.Yamada, and N.Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 363–371, New York, NY, USA, 2008. ACM.
11) K.R. Canini, L.Shi, and T.L. Griffiths. Online Inference of Topics with Latent Dirichlet Allocation. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.
12) T.P. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft, 2000.
13) R.Neal and G.Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer, 1998.
14) M.A. Sato and S.Ishii. On-line EM Algorithm for the Normalized Gaussian Network. *Neural Computation*, 12(2):407–432, 2000.
15) P.Liang and D.Klein. Online EM for Unsupervised Models. In *North American Association for Computational Linguistics (NAACL)*, 2009.
16) H.Robbins and S.Monro. A stochastic approximation method. In *Annals of Mathematical Statistics*, pages 400–407, 1951.