2004－SLP－53　(1)
2004／10／22

混合ディリクレ分布パラメータの階層ベイズモデルを用いたスムージング法

貞光 九月 †　　待鳥 裕介 ††　　山本 幹雄 †

†筑波大学 システム情報工学研究科, {sadamitsu@milab.is,myama@cs}.tsukuba.ac.jp
††筑波大学 情報学類, machi@milab.is.tsukuba.ac.jp

概要　文脈／文書中の話題を利用した適応的言語モデルを検討した。すでに提案している混合ディリクレ分布をユニグラムモデルの事前分布としたモデルは低い混合数 (10〜20) で最高性能となり、高い混合数では過適応してしまうため性能が悪化する。本稿では、高い混合数でも過適応をある程度抑えるスムージング法を提案し、性能向上を試みる。スムージング法は混合ディリクレ分布のパラメータの事前分布を仮定した階層ベイズモデルを利用した。評価実験によって 500 混合くらいまでは混合数に応じて性能が向上することを示す。

A smoothing method for parameters of Dirichlet mixtures using hierarchical Bayesian models

Kugatsu SADAMITSU †,　Yuusuke MACHITORI ††　and　Mikio YAMAMOTO †

†Graduate School of Systems & Information Engineering, University of Tsukuba,
{sadamitsu@milab.is,myama@cs}.tsukuba.ac.jp
††Collage of Information Sciences, University of Tsukuba, machi@milab.is.tsukuba.ac.jp

Abstract　We have investigated an adaptive language model which is aware of topics of context or documents. In the previous paper, we showed the language model using Dirichlet mixture distribution as a prior of unigram models gained the lowest perplexity at the small number of mixtures. However, the overfitting problem aggravated the perplexity of the model at the large number of mixtures. In this paper, we propose a smoothing method for Dirichlet mixture models to partially solve the overfitting problem. We assume a prior of parameters of Dirichlet mixture, that is, hierarchical Bayesian models, and describe an estimation method for the parameters and hyperparameters. Experimental results show the perplexity of the new model declease monotonically along with the number of mixture.

1　はじめに

文書や文脈で述べられている話題や分野を単語出現確率の偏りとしてモデル化し、その話題を捉えることによって言語モデルの精度を向上させる方法は 10 年ほど前から盛んに研究されている [例えば (Kneser&Steinbiss 1993) や (Bellegarda 1998)]。最近ではアスペクト・モデルと呼ばれる一連の確率モデルにより、話題モデルの精密化とロバストなパラメータ推定が可能となっている。これを利用した言語モデルの性能は向上しており (Gildea&Hofmann 1999)(秋田&河原 2003) 同音異義語誤り検出などに応用されている (三品 他 2004)。具体的なアスペクト・モデルとしては PLSA(Probabilistic LSA)(Hofmann 1999) や LDA(Latent Dirichlet Allocation) (Blei&Jordan 2001) が有名である。これらは話題ごとにユニグラムモデル (多項分布) を用意しておき、確率的に混合することで話題をモデル化する手法である。

これに対して我々は、話題による単語出現確率の偏りの分布を混合ディリクレ分布で直接モデル化する方法を提案しており、パープレキシティの観点で、低い混合数で PLSA や LDA よりも高い性能であることを示した (山本 他 2003)。しかし、混合ディリクレ分布はユニトピックモデル (一つの文書に一つの話題しか仮定しない) であるため、過適応しやすく高い混合数で性能が悪化することが分かっている。

本稿では、混合ディリクレモデルのパラメータ推定における過適応の問題を緩和するために、階層ベイズモ

デルを仮定したスムージング法を提案・検討する。ベイズ学習で問題となる期待値計算のための積分近似には、reversing EM(Minka 2001) を利用した。以下、2 節で基本的な混合ディリクレモデルの概要、3 節で過適応の問題、4 節で提案手法のスムージング法について述べ、5 節でパープレキシティの実験によってこれまで過適応していた高い混合数のモデルでも性能が向上することを示す。

2 混合ディリクレモデルとパラメータ推定

本節では混合ディリクレモデルと従来のパラメータ推定法について簡単に紹介する。詳しくは (山本 他 2003) を参照のこと。

V 次元の単体 $\Delta(V)$ 上の確率変数 $\boldsymbol{p} = (p_1, p_2, ..., p_V)$ に対して、M 個のディリクレ分布 $P_{Dir}(\boldsymbol{p}; \boldsymbol{\alpha}_m)$ を $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_M)$ で重み付けした混合ディリクレ分布 $P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M)$ は次のように定義される。

$$P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) = \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\prod_{v=1}^{V} \Gamma(\alpha_{mv})} \prod_{v=1}^{V} p_v^{\alpha_{mv}-1}$$

ここで、M は混合数、$\boldsymbol{\alpha}_1^M = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_M)$、$\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, ..., \alpha_{mV})$ は第 m コンポーネントのディリクレ分布のパラメータ、$\alpha_m = \sum_v \alpha_{mv}$ である。

文書中の単語出現頻度からパラメータを推定するために、混合ディリクレ分布と多項分布の合成分布である混合 Polya 分布をゆう度関数として最ゆう推定を行う。$\boldsymbol{y}_i$ をある文脈または文書、$y_{iv}(v = 1, 2, ..., V)$ を $\boldsymbol{y}_i$ に出現する各単語の出現頻度、$P_{Mul}(\boldsymbol{y}_i|\boldsymbol{p})$ を $\boldsymbol{p} \in \Delta(V)$ がパラメータである多項分布とすると、混合 Polya 分布 $P_{PM}(\boldsymbol{y}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M)$ は次のように定義される。

$$\begin{aligned} P_{PM}(\boldsymbol{y}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) &= \sum_{m=1}^{M} \lambda_m P_{Polya}(\boldsymbol{y}_i; \boldsymbol{\alpha}_1^M) \\ &= \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + y_i)} \prod_{v=1}^{V} \frac{\Gamma(y_{iv} + \alpha_{mv})}{\Gamma(\alpha_{mv})} \end{aligned}$$

ここで、$y_i = \sum_v y_{iv}$、$\alpha_m = \sum_v \alpha_{mv}$ である。$\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_D)$ とすると、対数ゆう度関数 $\mathcal{L}(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M)$ は混合 Polya 分布の積となり、これを最大化するために EM アルゴリズムを用いると以下のようなパラメータ $\boldsymbol{\lambda}$ と $\boldsymbol{\alpha}$ の更新式を得る。$\bar{\boldsymbol{\lambda}}$,$\bar{\boldsymbol{\alpha}}$ は現在の値である。

$$\lambda_m \propto \sum_i P_{im} \tag{1}$$

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_i P_{im}\{\Psi(y_{iv} + \bar{\alpha}_{mv}) - \Psi(\bar{\alpha}_{mv})\}}{\sum_i P_{im}\{\Psi(y_i + \bar{\alpha}_m) - \Psi(\bar{\alpha}_m)\}} \tag{2}$$

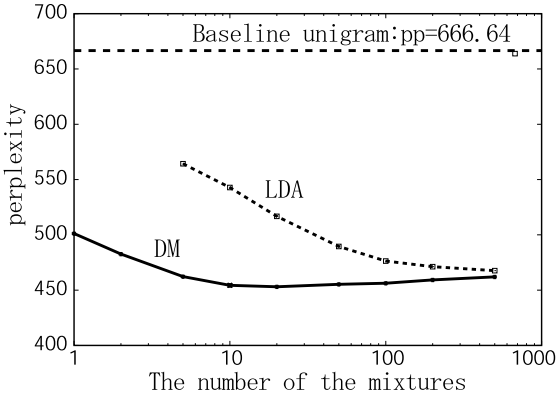

図 1: ディリクレモデルの過適応

ここで $\Psi(x)$ は digamma 関数と呼ばれる対数ガンマ関数の 1 階微分である。また、P_{im} は各文書のトピックを表す隠れ変数 z_i の事後確率であり、以下のように定義される。

$$P_{im} = P(z_i = m|\boldsymbol{y}_i; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}_1^M) = \frac{\bar{\lambda}_m P_{Polya}(\boldsymbol{y}_i; \bar{\boldsymbol{\alpha}}_m)}{\sum_{m'} \bar{\lambda}_{m'} P_{Polya}(\boldsymbol{y}_i; \bar{\boldsymbol{\alpha}}_{m'})}$$

leaving-one-out 法を用いるとより高速な更新式が得られる (山本 他 2003)。

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_i P_{im}\{y_{iv}/(y_{iv} - 1 + \bar{\alpha}_{mv})\}}{\sum_i P_{im}\{y_i/(y_i - 1 + \bar{\alpha}_m)\}} \tag{3}$$

3 混合ディリクレモデルの過適応

混合ディリクレモデルおよびアスペクトモデル (LDA)(Blei&Jordan 2001) の混合数に対するテストセット・パープレキシティの関係を図 1 に示す。パープレキシティは、文書中の 20 単語毎に単語履歴から予測分布を求めて計算した (山本 他 2003)。図 1 が示すように、混合ディリクレモデル (図中'DM') は低い混合数 (20 混合) で LDA よりも低いパープレキシティを達成するが、混合数が高くなると性能が悪化することが分かる。これは過適応の結果であるが、本節ではそのメカニズムを考察し、過適応を緩和する方法を検討する。

混合ディリクレモデルは 1 記事 1 トピックのユニトピックモデルであるため、前節で述べた学習アルゴリズム中の P_{im}(学習データ中のある i 番目の記事のカウントを各トピックに分配する割合) が極端な値となる。ほとんどの場合、あるトピック m に対して $P_{im} \approx 1$ で、他のトピック m' に対しては $P_{im'} \approx 0$ となる。これは、全学習記事を M 個のトピックにまず分割 (ハードクラスタリング) した後、各分割されたデータをもとにそれぞれのディリクレ分布を学習するのとほぼ等価である。LDA のようなマルチトピックモデルでは、ある記事が

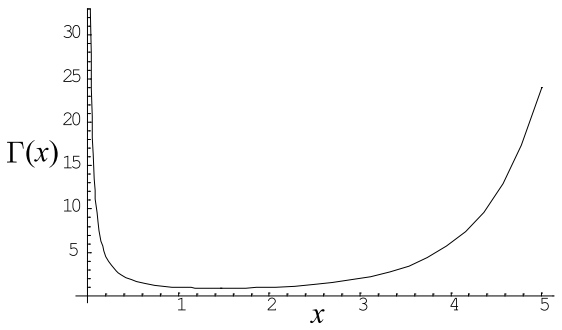

図 2: ガンマ関数

まるごとある特定のトピックにしか関係しないような分割は行われない。このように、混合ディリクレモデルでは学習データをトピック数 M 個に分割してしまうため、各トピックのディリクレ分布パラメータを推定するデータ量が減ってしまう。この結果、あるトピックでは出にくい単語がまったく現れない可能性が大きくなり、基本的に最ゆう推定である前節の推定方法ではデータ中に出現しない単語に対応するパラメータ α が極端に小さくなってしまう。このような状況は混合数が多いときに顕著である。たとえば、図 1 の実験で用いた推定値を観察すると、10^{-5} 以下の α_{mv} は 2 万単語中、10 混合で 2-3 割、100 混合で 8-9 割に達する。

小さな α を持つ単語が 1 回以上出現すると $\{\Gamma(y_{iv}+\alpha_{mv})\}/\{\Gamma(\alpha_{mv})\}$ の項によって Polya 分布の確率は小さくなる。これは図 2 のようなガンマ関数の特性より、小さな α_{mv} に対して分母が非常に大きくなるためである。該当する単語が出現しなければ (すなわち、$y_{iv}=0$)、分母・分子が同じになりキャンセルする。このように、小さな α_{mv} に対応する単語 v が数個出現すると、Polya 分布の確率が大きく減少するが、混合数が大きくなると、すべてのトピックにおいてもこれが生じやすくなる。

結局、$n$gram モデルのときと同じであるが、解決策としては出現しない単語に対応する α_{mv} に対しても極端に小さな値を割り付けないようにすればよい。

4 階層ベイズモデルを用いたスムージング

4.1 モデルパラメータの分割と事前分布

本節では、ディリクレ分布のパラメータに事前分布を仮定し、ベイズ学習を用いることで過適応を緩和する方法を述べる。まず、$\boldsymbol{\alpha}_m$ を、$s_m=\sum_v \alpha_{mv}$ と $\boldsymbol{r}_m=r_{m1},r_{m2},...,r_{mV}$ $(r_{mv}=\alpha_{mv}/s_m)$ に分解する (Minka 2003)。このとき、$\boldsymbol{r}_m$ は m 番目のディリクレ分布の確率変数 $\boldsymbol{p}$ の期待値であり、s_m はディリクレ分布の期待値からのばらつき度を表す。s_m が大きいほど、期待値の近辺に確率が集中する。α_{mv} の相対的な大きさは r_{mv} にのみ反映されるので、$\boldsymbol{r}$ にのみ事前分布を仮定する。ここで $\boldsymbol{r}_m$ は、その定義上必ず単体上に位置するため、事前分布として $\boldsymbol{r}_m$ を確率変数とするディリクレ分布を導入する。過適応の緩和が目的であるため、以下の式で示すように各ディリクレ分布のパラメータ (ハイパーパラメータ) はすべて等しく β とおく。

$$P(\boldsymbol{r};\beta)=\prod_{m=1}^{M} P_{Dir}(\boldsymbol{r}_m;\beta)$$

以上で述べたモデルのグラフィカルモデルを、図 3(右) に示す。図中の大きな 2 つの四角は、外側が D 個の文脈／文書の集合、内側が各文脈／文書ごとの L 個の単語を表現する。丸は確率変数、二重丸はモデルパラメータである。w は単語、矢印は変数 (データ) 間の依存関係 (条件付確率) を表す (矢の羽側が条件)。

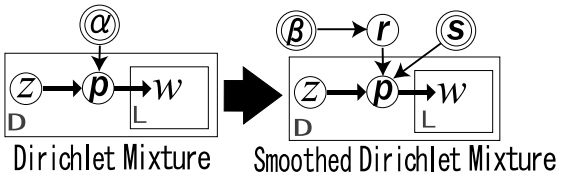

図 3: スムージングモデルのグラフィカルモデル表現

4.2 パラメータ推定

スムージングした混合ディリクレ分布のパラメータ $\boldsymbol{\lambda},\boldsymbol{s}$ と、ハイパーパラメータ β については最ゆう推定で求める。確率変数である $\boldsymbol{r}$ については、事後分布を求めてから期待値をとることによって決定する。実際にはそれぞれの値を交互に更新する繰り返しアルゴリズムを使用する。推定アルゴリズムの導出は比較的単純であるが、ゆう度と期待値の計算時に $\boldsymbol{r}$ の事後分布の積分が必要なため、事後分布を積分可能な式で近似する必要がある。近似手法にはさまざまなものがあるが、本稿では比較的単純である reversing EM(Minka 2001) によって近似した。

パラメータとハイパーパラメータの推定に必要なゆう度関数は以下のようになる。

$$\begin{aligned}P(\boldsymbol{y};\boldsymbol{s},\boldsymbol{\lambda},\beta)&=\int P(\boldsymbol{y},\boldsymbol{r};\boldsymbol{s},\boldsymbol{\lambda},\beta)d\boldsymbol{r}\\&=\int\left\{\prod_i P_{PM}(\boldsymbol{y}_i|\boldsymbol{r};\boldsymbol{\lambda},\boldsymbol{s})\right\}P(\boldsymbol{r};\beta)d\boldsymbol{r}\\&=\int\left\{\prod_i\sum_m \lambda_m P_{Polya}(\boldsymbol{y}_i|\boldsymbol{r}_m;s_m)\right\}P(\boldsymbol{r};\beta)d\boldsymbol{r}\end{aligned}$$

混合モデルであるので、2 節で定義した P_{im} を利用して、EM アルゴリズムの下限を導入する。ここで、

$\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{r}},\bar{\boldsymbol{s}},\bar{\beta}$ は各パラメータの現在値であるとする。

$$
\begin{aligned}
&P(\boldsymbol{y};\boldsymbol{\lambda},\boldsymbol{s},\beta)\\
&\geq \int\left[\prod_i\prod_m\left\{\frac{\lambda_m P_{Polya}(\boldsymbol{y}_i|\boldsymbol{r}_m;s_m)}{P_{im}}\right\}^{P_{im}}\right]P(\boldsymbol{r};\beta)d\boldsymbol{r}\\
&=\prod_m\left[\left\{\prod_i\left(\frac{\lambda}{P_{im}}\right)^{P_{im}}\right\}\right.\\
&\quad\left.\times\int\left\{\prod_i P_{Polya}^{P_{im}}(\boldsymbol{y}_i|\boldsymbol{r}_m;s_m)\right\}P(\boldsymbol{r}_m;\beta)d\boldsymbol{r}_m\right]
\end{aligned}
$$

さらに、積分の部分に各分布式を代入して整理すると以下のようになる。

$$
\begin{aligned}
&\int\left\{\prod_i P_{Polya}^{P_{im}}(\boldsymbol{y}_i|\boldsymbol{r}_m;\boldsymbol{s})\right\}P_{Dir}(\boldsymbol{r}_m;\beta)d\boldsymbol{r}_m\\
&=\left\{\prod_i\frac{\Gamma^{P_{im}}(s_m)}{\Gamma^{P_{im}}(s_m+y_i)}\right\}\frac{\Gamma(V\beta)}{\Gamma^V(\beta)}\\
&\quad\times\int\prod_v\left\{\prod_i\frac{\Gamma^{P_{im}}(s_m r_{mv}+y_{iv})}{\Gamma^{P_{im}}(s_m r_{mv})}\right\}r_{mv}^{\beta-1}d\boldsymbol{r}_m
\end{aligned}
$$

このままでは上記の積分は困難であるので、一部の式に下限の関数を導入し、近似を行なった後に積分を行なう。下限の関数は、ガンマ関数の比になっている部分について文献 (Minka 2003) の (127) 式を用いた。

$$
\begin{aligned}
&\int\prod_v\left\{\prod_i\frac{\Gamma^{P_{im}}(s_m r_{mv}+y_{iv})}{\Gamma^{P_{im}}(s_m r_{mv})}\right\}r_{mv}^{\beta-1}d\boldsymbol{r}_m\\
&\geq\int\prod_v\left[\left\{\prod_i c_{imv}^{P_{im}}(s_m r_{mv})^{P_{im}a_{imv}}\right\}r_{mv}^{\beta-1}\right]d\boldsymbol{r}_m\\
&=\left\{\prod_i\prod_v c_{imv}^{P_{im}}s_m^{P_{im}a_{imv}}\right\}\frac{\prod_v\Gamma(\sum_i P_{im}a_{imv}+\beta)}{\Gamma(\sum_v\sum_i P_{im}a_{imv}+V\beta)}
\end{aligned}
$$

ここで、

$$
\begin{aligned}
a_{imv}&=\{\Psi(\bar{s}_m\bar{r}_{mv}+y_{iv})-\Psi(\bar{s}_m\bar{r}_{mv})\}\bar{s}_m\bar{r}_{mv},\\
c_{imv}&=\frac{\Gamma(\bar{s}_m\bar{r}_{mv}+y_{iv})}{\Gamma(\bar{s}_m\bar{r}_{mv})}(\bar{s}_m\bar{r}_{mv})^{-a_{imv}}.
\end{aligned}
$$

近似精度を高めるために、積分結果であるゆう度を最大とする $\bar{\boldsymbol{r}}$ を求めるのが reversing EM (Minka 2001) である。積分結果が最大であるということは、事後分布の下限による近似がもっとも真の分布に近づいていると解釈できる。$\bar{\boldsymbol{r}}$ はたくさんの箇所に出現しているので、このまま最大化することは困難であるが、幸いなことに「ゆう度最大とする $\bar{\boldsymbol{r}}$」と「$\boldsymbol{r}$ の事後分布の確率を最大とする $\boldsymbol{r}$」が近似的に一致する[1]。よって、最適な近似は事後分布 $P(\boldsymbol{r}|\boldsymbol{y};\boldsymbol{\lambda},\boldsymbol{s},\beta)$ が最大となる $\boldsymbol{r}$ を求めることによって得られる (すなわち $\boldsymbol{r}$ の MAP 推定)。$\boldsymbol{r}$ の事後分布は $\boldsymbol{y}$ との同時分布に比例するので、最大化すべき対数ゆう度 $\mathcal{L}(\boldsymbol{r})$ は次のようになる。

$$\mathcal{L}(\boldsymbol{r})=\log P(\boldsymbol{y},\boldsymbol{r};\boldsymbol{\lambda},\boldsymbol{s},\beta)$$

この式に EM アルゴリズムの下限と、文献 (Minka 2003) の (127) 式を用いることによって、以下のように下限の関数を得る。

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{r})=\log\left[\left\{\prod_i\sum_m\lambda_m P_{Polya}(\boldsymbol{y}_i|\boldsymbol{r}_m)\right\}\left\{\prod_m P_{Dir}(\boldsymbol{r}_m;\beta)\right\}\right]\\
&\geq\log\left[\prod_i\prod_m\left\{\frac{\lambda_m P_{Polya}(\boldsymbol{y}_i|\boldsymbol{r}_m)}{P_{im}}\right\}^{P_{im}}\left\{\prod_v P_{Dir}(\boldsymbol{r}_m;\beta)\right\}\right]\\
&=\sum_m\left[\left\{\sum_i P_{im}\sum_v\log\frac{\Gamma(y_{iv}+s_m r_{mv})}{\Gamma(s_m r_{mv})}\right\}\right.\\
&\qquad\left.+\sum_v(\beta-1)\log r_{mv}\right]+const.\\
&\geq\sum_m\left\{\sum_i P_{im}\sum_v\log c_{imv}(s_m r_{mv})^{a_{imv}}\right.\\
&\qquad\left.+\sum_v(\beta-1)\log r_{mv}\right\}+const.\\
&=\sum_m\left[\sum_v\left\{\left(\sum_i P_{im}a_{imv}\right)+\beta-1\right\}\log r_{mv}\right]+const.
\end{aligned}
$$

最後の式を微分して 0 とおいて、以下のような r_{mv} の更新式が得られる。

$$r_{mv}\propto\sum_i a_{imv}+\beta-1 \tag{4}$$

ゆう度を最適に近似できたところで、残りのパラメータについてゆう度関数を最大とする値を求める。$\boldsymbol{\lambda}$ と $\boldsymbol{s}$ の対数ゆう度関数はそれぞれ以下のようになる (関係のない部分は省略)。

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\lambda})&=\sum_m\left\{\sum_i P_{im}\right\}\log\lambda_m\\
\mathcal{L}(\boldsymbol{s})&=\sum_{m,i}P_{im}\log\left\{\frac{\Gamma(s_m)}{\Gamma(s_m+y_i)}+\sum_v a_{imv}\log s_m\right\}
\end{aligned}
$$

それぞれの対数ゆう度関数を最大化する fixed-point iteration による更新式は以下のようになる。

$$\lambda_m\propto\sum_i P_{im} \tag{5}$$

$$s_m=\frac{\sum_i P_{im}\sum_v a_{imv}}{\sum_i P_{im}\{\Psi(y_i+\bar{s}_m)-\Psi(\bar{s}_m)\}} \tag{6}$$

β の推定には、対数をとらないゆう度 $f(\beta)$ の下限を設定し、その下限を最大化することにより fixed-point

[1] 紙面の都合で証明は省略する。P_{im} を一定と仮定し、$\boldsymbol{r}$ の事後分布を微分した関数をゼロとおいた方程式を、ゆう度を微分した式に代入するとこれまたゼロとなることによって証明できる。別の機会に報告する予定である。

iteration の式を得る。下限は、文献 (Minka 2003) の (125) 式を用いた。

$$
\begin{aligned}
f(\beta) &= \prod_m \frac{\Gamma(V\beta)}{\Gamma^V(\beta)} \frac{\prod_v \Gamma(\sum_i P_{im} a_{imv} + \beta)}{\Gamma(\sum_v \sum_i P_{im} a_{imv} + V\beta)} \\
&= \prod_m \frac{\Gamma(V\beta)}{\Gamma(\sum_v Pa_{mv} + V\beta)} \prod_v \frac{\Gamma(Pa_{mv} + \beta)}{\Gamma(\beta)} \\
&\geq \prod_m \frac{\Gamma(V\bar{\beta}) \exp\{(V\bar{\beta}) b'_m\}}{\Gamma(V\bar{\beta} + \sum_v Pa_{mv})} \prod_v c'_{mv} \beta^{a'_{mv}} \\
&= g(\beta)
\end{aligned}
$$

ここで、

$$
\begin{aligned}
Pa_{mv} &= \sum_i P_{im} a_{imv}, \\
a'_{mv} &= \{\Psi(\bar{\beta} + Pa_{mv}) - \Psi(\bar{\beta})\}\bar{\beta}, \\
b'_m &= \Psi\left(V\bar{\beta} + \sum_v Pa_{mv}\right) - \Psi(V\bar{\beta}), \\
c'_{mv} &= \frac{\Gamma(Pa_{mv} + \bar{\beta})}{\Gamma(\bar{\beta})} \bar{\beta}^{-a'_{mv}}.
\end{aligned}
$$

$g(\beta)$ を最大化することによって β の更新式を得る。

$$
\beta = \frac{\sum_m \sum_v a'_{mv}}{V \sum_m b'_m} \tag{7}
$$

最後に $\boldsymbol{r}$ の推定方法を述べる。これまで、他パラメータの最ゆう推定値を得るために、$\boldsymbol{r}$ の事後分布の下限による近似をその積分 (ゆう度) が最大になるように最適化した。これは事後分布のよい近似にもなっているので、これを利用して期待値をとることによって $\boldsymbol{r}$ の推定値を得る。事後分布は以下のように近似されている。

$$
P(\boldsymbol{r}|\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{s}, \beta) \propto \prod_{mv} r_{mv}^{(\sum_i P_{im} a_{imv}) + \beta - 1}
$$

r_{mv} の期待値は次のように計算できる。

$$
\begin{aligned}
E[r_{mv}]_{P(\boldsymbol{r}|\boldsymbol{y};\boldsymbol{\lambda},\boldsymbol{s},\beta)} &= \int r_{mv} P(\boldsymbol{r}|\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{s}) d\boldsymbol{r} \\
&= \frac{\int r_{mv} \prod_{m',v'} r_{m'v'}^{(\sum_i P_{im} a_{im'v'}) + \beta - 1} d\boldsymbol{r}}{\int \prod_{m',v'} r_{m'v'}^{(\sum_i P_{im} a_{im'v'}) + \beta - 1} d\boldsymbol{r}} \\
&= \frac{\sum_i (P_{im} a_{imv} + \beta)}{\sum_{v'} \sum_i (P_{im} a_{imv'} + \beta)}
\end{aligned} \tag{8}
$$

上式の期待値計算はすべての他のパラメータ ($\bar{\boldsymbol{r}}$ を含む) が収束した後、最後に一度だけ計算する。

5 実験

文脈に対する動的適応を用いたテストセットパープレキシティを計算し、スムージングを行った場合と行っていない場合の混合ディリクレ分布を比較した。

学習データは 1999 年版毎日新聞記事 98211 記事、テストセットは 1998 年版毎日新聞から 40 単語以上を含む 495 記事をランダムに選択した。語彙は学習データ中の高頻度 20000 語とした (カバー率 97.1%)。

学習法は、まずスムージングされていないパラメータを求めた後、そのパラメータを初期値としてスムージング法による学習を行った。これは学習の高速化のためである。スムージングされていないパラメータは、leaving-one-out 法を用い、学習の終了条件はパープレキシティの減少率が 0.1%以下になった時点とした。スムージング有りの場合には、まず $\boldsymbol{\lambda}$ と $\boldsymbol{s}$ について、それぞれパープレキシティ減少率が 0.1%以下になるまで繰り返し計算する。$\boldsymbol{r}$ については、(4) 式を用いて事後分布の近似を最適とした後に、(8) 式で期待値を得るのが理想的であるが、実際に計算してみたところ、(4) 式において r がマイナスになることがまれに生じるため、はじめから (8) 式を更新式に用いることで、MAP 推定の近似とし、パープレキシティ減少率が 0.1%以下になるまで繰り返し計算した。β については、β 自身の値の変化率が 0.1%以下になった時点で収束と判定した。この $\boldsymbol{\lambda} \rightarrow \boldsymbol{s} \rightarrow \boldsymbol{r} \rightarrow \beta$ のパラメータ推定の流れを 1 セットとし、6 回繰り返してパラメータを得た。さらに、今回得られた β の値が妥当なものであるかを評価するため、β を固定した場合のスムージングも行った。用いた値は、$0.01, 0.1, 1, 10$ の 4 つである。混合数は 5,10,20,50,100,200,500 のモデルを作成した。学習時間は、XEON 2.4GHz の Linux マシンを用いた場合、混合数 500 のモデルで leaving-one-out 法のみを用いた場合、5 日程度の時間を要し、スムージング有りの 500 混合の再学習では、プラス 2 日程度 (β 推定有り) の時間を要した。

ユニグラムパープレキシティの計算方法として、文脈が 20 単語増える毎にそこまでの全単語を用いて適応を行ない、次の 20 単語の確率を予測することを繰り返すヒストリ適応確率を計算した。未知語の確率はパープレキシティの計算から除いている。単純な ユニグラムモデルのテストセットパープレキシティは 666.64 である。

図 4 が実験の結果である。'estimate' とマークしているグラフが β を推測しながら行ったスムージング、数字のみマークしているものが、β を固定した場合のスムージングで、それぞれの数字は固定した β の値である。また、$\beta = 0$ の場合は、スムージングしていない場合と等価になる。つまり、$\boldsymbol{\alpha}$ の学習を最ゆう推定で行ったことと同じとなる (ただし、leaving-one-out 法を用いて学習したパラメータを初期値としているので、厳密に同じではない)。

β を推定によって求める手法は、スムージングを行わ

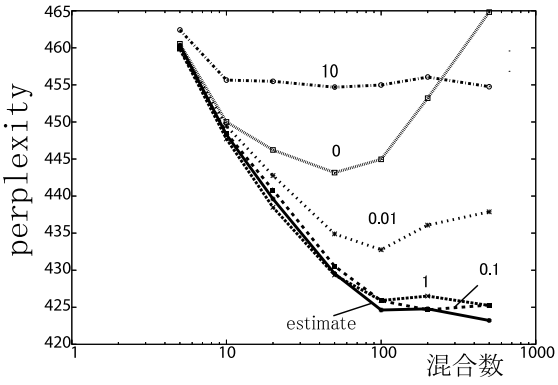

図 4: ヒストリ適応確率によるテストセットパープレキシティの比較

表 1: 各モデルのパープレキシティ最低値

Model	Perplexity
smoothed DM(β estimate)	423.21(36.5%)
smoothed DM($\beta = 0.01$)	432.76(35.1%)
smoothed DM($\beta = 0.1$)	424.65(36.3%)
smoothed DM($\beta = 1$)	425.23(36.2%)
smoothed DM($\beta = 10$)	454.69(31.8%)
smoothed DM($\beta = 0$)	443.17(33.5%)
LOO DM	453.06(32.0%)
LDA	467.61(29.9%)

ない場合より優れた結果を示しただけでなく、他の β を固定したものよりも良い結果を示したことから、4.2 節の推定方法の正当性が裏付けされているといえる。なお、各混合数における β の値の相違はほとんどみられず、0.16 付近に収束した (最低値は 100 混合の 0.162、最大値は 5 混合の 0.173)。

なお、それぞれの実験における結果のうち、もっとも低いパープレキシティを示した場合について、表 1 にまとめた。'smoothed DM' がスムージングを行った混合ディリクレモデル (括弧内は β の値)、'LOO DM' が、図 1 に示した leaving-one-out 単独学習での結果、'LDA' は同じく図 1 中に示した LDA の値である。また、表中パープレキシティ欄の括弧内の数字は、ベースラインからのパープレキシティ削減率を示す。

6 おわりに

混合ディリクレ分布のパラメータに事前分布を導入し、階層ベイズモデルでスムージングされたパラメータを学習する方法を提案・検討した。ベイズ学習の複雑な期待値計算には、reversing EM 法を用いることにより効率的な推定法を得ることができた。実験により、100 混合以上でも性能が劣化しないことを確認した。

参考文献

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2001). "Latent Dirichlet Allocation." In *Neural Information Processing Systems*, Vol. 14.

Gildea, D. and Hofmann, T. (1999). "Topic-based language models using em." In *Proc. of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2167–2170.

Hofmann, T. (1999). "Probabilistic Latent Semantic Indexing." In *Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pp. 50–57 Berkeley, California.

J.R.Bellegarda (1998). "A multispan language modeling framework for large vocabulary speech recognition." *IEEE Trans. Speech Audio Process*, **6** (5), pp. 456–467.

Minka, T. (2001). "Using lower bounds to approximate." http://www.stat.cmu.edu/ minka/papers/rem.html.

Minka, T. (2003). "Estimating a Dirichlet distribution." http://www.stat.cmu.edu/ minka/papers/dirichlet.html.

R.Kneser&V.Steinbiss (1993). "On the dynamic adaptation of stochastic language models." In *ICASP-93*, pp. II–586–589.

山本幹雄, 貞光九月, 三品拓也 (2003). "混合ディレクレ分布を用いた文脈のモデル化と言語モデルへの応用." 情報処理学会研究報告, SLP-48, pp. 29–34.

三品拓也, 貞光九月, 山本幹雄 (2004). "確率的 LSA を用いた日本語同音異義語誤りの検出・訂正." 情報処理学会論文誌, **45** (9), pp. 2168–2176.

秋田祐哉 河原達也 (2003). "話題と話者に関する PLSA に基づく言語モデル適応." 情報処理学会研究報告, SLP-49, pp. 67–72.