ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 26 № 3 С. 56–68 2020

УДК 519.853.4

СХОДИМОСТЬ АЛГОРИТМА АДДИТИВНОЙ РЕГУЛЯРИЗАЦИИ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ[1]

И. А. Ирхин, К. В. Воронцов

Задача вероятностного тематического моделирования заключается в следующем. По заданной коллекции текстовых документов требуется найти условное распределение каждого документа по темам и условное распределение каждой темы по словам (или термам). Для решения данной задачи используется принцип максимума правдоподобия. Задача имеет в общем случае бесконечное множество решений, то есть является некорректно поставленной по Адамару. В рамках подхода ARTM — аддитивной регуляризации тематических моделей к основному критерию добавляется взвешенная сумма нескольких дополнительных критериев регуляризации. Численный метод для решения данной задачи является разновидностью итерационного EM-алгоритма, который выписывается в общем виде для произвольного гладкого регуляризатора, в том числе и для линейной комбинации гладких регуляризаторов. В работе исследуется вопрос о сходимости данного итерационного процесса. Получены достаточные условия сходимости, при которых процесс сходится к стационарной точке регуляризованного логарифма правдоподобия. Полученные ограничения на регуляризатор оказались не слишком обременительными. В работе даны их интерпретации с точки зрения практической реализации алгоритма. Предложена модификация алгоритма, которая улучшает его сходимость без дополнительных затрат времени и памяти. В экспериментах на коллекции новостных текстов показано, что предложенная модификации позволяет не только ускорить сходимость, но и улучшить значение оптимизируемого критерия.

Ключевые слова: обработка текстов естественного языка; вероятностное тематическое моделирование; вероятностный латентный семантический анализ; PLSA; латентное размещение Дирихле; LDA; аддитивная регуляризация тематических моделей; ARTM; EM-алгоритм; достаточные условия сходимости.

The problem of probabilistic topic modeling is as follows. Given a collection of text documents, find the conditional distribution over topics for each document and the conditional distribution over words or terms for each topic. Log-likelihood maximization is used to solve this problem. The problem has generally an infinite set of solutions, being ill-posed according to Hadamard. In the framework of Additive Regularization of Topic Models (ARTM), a weighted sum of regularization criteria is added to the main log-likelihood criterion. The numerical method for solving this optimization problem is a kind of iterative EM-algorithm. In ARTM it is inferred in a quite general form for an arbitrary smooth regularizer, as well as for a linear combination of smooth regularizers. This paper studies the problem of convergence of the EM iterative process. Sufficient conditions are obtained for the convergence to a stationary point of the regularized log-likelihood. The constraints imposed on the regularizer are not too restrictive. We give their interpretations from the point of view of the practical implementation of the algorithm. A modification of the algorithm is proposed that improves the convergence without additional time and memory costs. Experiments on the news text collection have shown that our modification both accelerates the convergence and improves the value of the criterion to which it converges.

Keywords: natural language processing; probabilistic topic modeling; probabilistic latent semantic analysis; PLSA; latent Dirichlet allocation; LDA; additive regularization of topic models; ARTM; EM-algorithm; sufficient conditions for convergence.

MSC: 90C30, 68T50.

[1]Работа выполнена в рамках проекта «Средства интеллектуального анализа больших массивов текстов», по Программе ЦК НТИ «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по договору МГУ им. М. В. Ломоносова с Фондом поддержки проектов НТИ от 15.08.2019 № 7/1251/2019. Работа также частично поддержана РФФИ, проект 20-07-00936.

Введение

Тематическое моделирование — одно из современных направлений *обработки естественного языка* (natural language processing, NLP). Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ, и какие термы образуют каждую тему. *Термами* могут быть слова, нормальные формы слов, словосочетания или термины, в зависимости от того, какие виды предварительной обработки текста были применены к данной коллекции. Тематическое моделирование не претендует на полноценное *понимание естественного языка* (natural language understanding, NLU), однако выявление тематики текстов можно считать определённым шагом в этом направлении.

Вероятностная тематическая модель (probabilistic topic model, PTM) описывает каждый документ дискретным распределением вероятностей на множестве тем, каждую тему — дискретным распределением вероятностей на множестве термов. Построенная модель позволяет преобразовать любой текст в вектор вероятностей тем. Важным преимуществом тематического векторного представления текста является его интерпретируемость. Каждая координата вектора показывает долю соответствующей темы в тексте, при этом семантика темы описывается частотным словарём термов, то есть фактически словами естественного языка.

Тематическое моделирование, как и *кластеризация документов*, относится к методам обучения без учителя и не требует какой-либо разметки текстов или экспертных оценок. Отличие в том, что при кластеризации документ целиком относится к одному кластеру, тогда как тематическая модель осуществляет *мягкую кластеризацию* (soft clustering), разделяя документ между несколькими кластерами-темами. Тематические модели называют также моделями мягкой би-кластеризации, поскольку термы также кластеризуются по темам. Это позволяет обходить проблемы синонимии и полисемии слов. Синонимы, употребляемые в схожих контекстах, группируются в одних и тех же темах. Многозначные слова и омонимы, наоборот, распределяют свои вероятности по нескольким семантически не связанным темам.

Перечисленные особенности вероятностного тематического моделирования делают его важным инструментом семантического анализа больших текстовых коллекций.

Построение тематической модели по коллекции документов является некорректно поставленной оптимизационной задачей приближённого стохастического матричного разложения, которая в общем случае имеет бесконечное множество решений. Согласно теории регуляризации А. Н. Тихонова [9], решение такой задачи возможно доопределить и сделать устойчивым. Для этого к оптимизационному критерию добавляется *регуляризатор* — дополнительный критерий, учитывающий специфические особенности прикладной задачи или знания предметной области. В сложных приложениях дополнительных критериев может быть несколько.

Аддитивная регуляризация тематических моделей (additive regularization of topic models, ARTM) — это многокритериальный подход, в котором для оптимизации параметров модели используется взвешенная сумма критериев [11;14;15]. ARTM позволяет строить модели с требуемыми свойствами, суммируя регуляризаторы, исходно предлагавшиеся в различных моделях, главным образом, в рамках байесовского обучения [6]. Однако в байесовском обучении не существует общего подхода к комбинированию регуляризаторов от разных моделей. В ARTM для обучения модели с произвольной линейной комбинацией регуляризаторов используется один и тот же *EM-подобный алгоритм* (EM-like algorithm), при этом для добавления нового регуляризатора достаточно знать его частные производные по параметрам модели. Это приводит к модульной технологии тематического моделирования, которая реализована в библиотеке с открытым кодом `BigARTM`, `http://bigartm.org` [4; 12].

Подчеркнём, что ARTM не является ещё одной частной тематической моделью или методом — это общий подход к построению и комбинированию тематических моделей.

До сих пор в теории ARTM оставались открытыми вопросы о сходимости EM-алгоритма и о влиянии регуляризаторов на сходимость. В данной работе показано, что в ARTM итерации EM-алгоритма возможно интерпретировать как итерации обобщённого EM-алгоритма (Generalized EM, GEM) [3], для которого условия сходимости хорошо изучены [17]. В работе получены ограничения на регуляризаторы, обеспечивающие сходимость, и предложена модификация EM-алгоритма, улучшающая его сходимость.

1. Задача тематического моделирования с аддитивной регуляризацией

Пусть D — конечное множество (коллекция) текстовых документов, W — конечное множество (словарь) всех употребляемых в них термов, T — конечное множество тем. Каждый документ $d \in D$ представляет собой последовательность n_d термов $(w_1, \ldots, w_{n_d})$ из словаря W. Примем гипотезу «мешка слов», согласно которой порядок термов в документе не важен. Обозначим через n_{dw} число вхождений терма w в документ d.

Коллекцию документов будем рассматривать как множество троек $(d, w, t) \in D \times W \times T$, выбранных случайно и независимо из дискретного распределения $p(d, w, t)$. При этом документы d и термы w являются наблюдаемыми переменными, темы t являются латентными (скрытыми) переменными.

Примем гипотезу условной независимости $p(w\,|\,d, t) = p(w\,|\,t)$, согласно которой распределение термов в теме одинаково для всех документов. Тогда, по формуле полной вероятности,

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,d,t)p(t\,|\,d) = \sum_{t\in T} p(w\,|\,t)p(t\,|\,d) = \sum_{t\in T} \phi_{wt}\theta_{td},$$

где $\phi_{wt} = p(w\,|\,t)$ — неизвестное распределение термов в темах, $\theta_{td} = p(t\,|\,d)$ — неизвестное распределения тем в документах. Условная вероятность $p(w\,|\,d)$ называется вероятностной тематической моделью документа, переменные ϕ_{wt} и θ_{td} являются параметрами этой модели.

Задача вероятностного тематического моделирования заключается в том, чтобы найти параметры модели по эмпирическим данным n_{dw}. Для этого решается задача максимизации логарифма правдоподобия

$$L(\Phi, \Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w\,|\,d) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta} \tag{1.1}$$

при ограничениях неотрицательности и нормировки:

$$\phi_{wt} \geq 0, \quad \sum_{w\in W} \phi_{wt} = 1, \qquad \theta_{td} \geq 0, \quad \sum_{t\in T} \theta_{td} = 1,$$

где Φ и Θ — матрицы параметров ϕ_{wt} и θ_{td} соответственно.

Задача (1.1) является некорректно поставленной задачей приближённого стохастического матричного разложения $(\frac{n_{dw}}{n_d}) \approx \Phi\Theta$, имеющей в общем случае бесконечное множество решений. Чтобы выбрать из него наиболее подходящее решение, вводятся дополнительные критерии — регуляризаторы $R_i(\Phi, \Theta) \to \max$, $i = 1, \ldots, k$. В подходе ARTM [11; 13; 14] предлагается максимизировать взвешенную сумму всех регуляризаторов $R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$ совместно с основным критерием правдоподобия:

$$L(\Phi, \Theta) + R(\Phi, \Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \log \sum_{t\in T} \phi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}, \tag{1.2}$$

при тех же ограничениях неотрицательности и нормировки.

Наиболее известные тематические модели PLSA и LDA являются частными случаями регуляризации. В модели вероятностного латентного семантического анализа PLSA [5] регуляризация не используется, $R(\Phi,\Theta)=0$. В модели латентного размещения Дирихле LDA [2] регуляризатором является логарифм правдоподобия априорного распределения Дирихле,

$$R(\Phi,\Theta)=\sum_{t\in T}\sum_{w\in W}(\beta_w-1)\ln\phi_{wt}+\sum_{d\in D}\sum_{t\in T}(\alpha_t-1)\ln\theta_{td}$$

с гиперпараметрами β_w, α_t, которые на практике обычно фиксируются, но могут и оптимизироваться с помощью специальных численных методов [16]. В данной работе проблемы оптимизации гиперпараметров в критериях регуляризации, а также выбора коэффициентов регуляризации τ_i, не рассматриваются.

Применение теоремы Каруша–Куна–Таккера позволяет выписать систему уравнений для стационарных точек оптимизационной задачи (1.2). Решение данной системы методом простых итераций приводит к EM-подобному алгоритму, в котором на каждой итерации чередуются два шага: E-шаг (expectation) и M-шаг (maximization).

На E-шаге вычисляются значения условных вероятностей $p_{tdw}=p(t\,|\,d,w)$ по текущим значениям параметров ϕ_{wt} и θ_{td}:

$$p_{tdw}=\frac{\varphi_{wt}\theta_{td}}{\sum_s\varphi_{ws}\theta_{sd}}.$$

Данное выражение совпадает с формулой Байеса, поскольку, в силу гипотезы условной независимости, $p(t\,|\,d,w)=\frac{p(w\,|\,t)\;p(t\,|\,d)}{p(w\,|\,d)}$.

На M-шаге по условным вероятностям тем p_{tdw} для каждого терма в каждом документе вычисляются новые приближения параметров ϕ_{wt} и θ_{td} и вспомогательные переменные n_{dwt}, n_{wt}, n_{td}, n_t, n_d, r_{wt}, r_{td}:

$$\begin{aligned}
n_{dwt}&=n_{dw}p_{tdw},\\
n_{wt}&=\sum_{d\in D}n_{dwt}, & n_{td}&=\sum_{w\in d}n_{dwt},\\
n_t&=\sum_{w\in W}n_{wt}, & n_d&=\sum_{t\in T}n_{td},\\
r_{wt}&=\phi_{wt}\frac{\partial R}{\partial\phi_{wt}}, & r_{td}&=\theta_{td}\frac{\partial R}{\partial\theta_{td}},\\
\phi_{wt}&=\operatorname*{norm}_{w\in W}(n_{wt}+r_{wt}), & \theta_{td}&=\operatorname*{norm}_{t\in T}(n_{td}+r_{td}),
\end{aligned}$$

где $\operatorname*{norm}_{i\in I}(x_i)=\frac{(x_i)_+}{\sum_{j\in I}(x_j)_+}$ — операция нормировки, которая переводит произвольный числовой вектор $(x_i\colon i\in I)$ в дискретное вероятностное распределение, операция $(x_i)_+=\max(x_i,0)$ называется положительной срезкой.

Вспомогательные переменные n_* интерпретируются как оценки счётчиков: n_{dwt} — число вхождений терма w в документ d, связанных с темой t; n_{td} — число всех термов в документе d, связанных с темой t; n_{wt} — число раз, когда терм w был связан с темой t, во всей коллекции; n_t — число термов, связанных с темой t, во всей коллекции; n_d совпадает с длиной документа d.

Вспомогательные переменные r_{wt} и r_{td} будем называть регуляризационными поправками. Заметим, что при $R=0$, то есть в модели PLSA, $r_{wt}=0$, $r_{td}=0$, $\phi_{wt}=\frac{n_{wt}}{n_t}$, $\theta_{td}=\frac{n_{td}}{n_d}$.

2. Теорема о сходимости EM-алгоритма в ARTM

Достаточные условия для сходимости обобщённого EM-алгоритма GEM алгоритма были сформулированы в [17]. Мы собираемся использовать те же методы доказательства, интерпре-

тируя итерации EM-алгоритма ARTM как итерации GEM алгоритма.

Объединяя несколько теорем из [17] и адаптируя обозначения, нетрудно получить теорему, с помощью которой удобно доказывать сходимость EM-алгоритма в ARTM.

Теорема 1. *Пусть* $\{(\Phi^k, \Theta^k)\}$ *— траектория итерационного процесса, сгенерированная правилом* $(\Phi^{k+1}, \Theta^{k+1}) = M(\Phi^k, \Theta^k)$*, где* M *— непрерывное преобразование пары стохастических матриц. Пусть функция* $F(\Phi, \Theta)$ *ограничена сверху и строго возрастает под действием* M *на* (Φ, Θ)*. Тогда все предельные точки траектории* (Φ^k, Θ^k) *являются стационарными точками* F*. Если также* $\|\phi_{wt}^k - \phi_{wt}^{k+1}\| \to 0$ *и* $\|\theta_{td}^k - \theta_{td}^{k+1}\| \to 0$*, а множество стационарных точек* F *дискретно, то* (Φ^k, Θ^k) *сходится к некоторой стационарной точке* F*.*

О п р е д е л е н и е 1. Регуляризатор R является *δ-регулярным*, если на итерациях EM-алгоритма $\forall t\ \exists w$: $n_{wt} + r_{wt} > \delta$ и $\forall d\ \exists t$: $n_{td} + r_{td} > \delta$. Если регуляризатор обладает свойством δ-регулярности при некотором $\delta > 0$, то будем говорить, что регуляризатор сильно регулярен; при $\delta = 0$ будем просто говорить, что он регулярен.

Регулярность гарантирует, что в операции norm не возникнет деления на нуль, то есть итерации корректно определены. Сильная же регулярность позволяет утверждать, что преобразования, которые производятся на итерациях алгоритма, являются непрерывными по (Φ, Θ). Это свойство легко выполняется на практике: если значение $n_{wt}+r_{wt}$ (или $n_{td}+r_{td}$) становится меньше δ, то вся тема (весь документ) исключается из модели и итерации продолжаются.

О п р е д е л е н и е 2. Регуляризатор R *сохраняет нуль*, если на итерациях алгоритма из $n_{wt} = 0$ следует $\phi_{wt} = 0$ и из $n_{td} = 0$ следует $\theta_{td} = 0$.

Это определение формализует следующие свойство итерационного процесса: если на какой-либо итерации значение ϕ_{wt} стало равным нулю, то оно будет оставаться нулевым на последующих итерациях, и аналогично для θ_{td}. Для регуляризатора данное свойство легко проверяется аналитически. На практике многие регуляризаторы им обладают. Регуляризатор модели LDA, вообще говоря, не обладает данным свойством при $\beta_w > 1$ или $\alpha_t > 1$, так как при $n_{wt} = 0$ вполне может оказаться, что $\phi_{wt} > 0$. Однако при использовании ненулевой инициализации ϕ_{wt} значение n_{wt} не может обратиться в нуль. Поэтому и для такого регуляризатора условие сохранения нуля выполняется.

О п р е д е л е н и е 3. Регуляризатор R называется *ϵ-разреживающим*, если на итерациях EM-алгоритма $\phi_{wt}, \theta_{td} \notin (0, \epsilon)$.

Некоторые регуляризаторы имеют неограниченную в окрестности нуля производную, поэтому при реализации EM-алгоритма параметры, меньшие некоторого ϵ, зануляются. Это приводит к тому, что значения в матрице параметров оказываются отделены от нуля. Именно эта особенность отражена в данном определении.

О п р е д е л е н и е 4. Регуляризатор R *корректный*, если на итерациях EM-алгоритма из $n_{dw} > 0$ следует $p_{tdw} > 0$ хотя бы для одной темы t.

Если модель даёт нулевую оценку вероятности $p(w\,|\,d) = 0$ при том, что терм w встречается в документе, $n_{dw} > 0$, то логарифм правдоподобия становится неограниченным, $L \to -\infty$. На практике этого легко избежать, если использовать регуляризатор сглаживания фоновых тем [13]. Он гарантирует, что для любого терма в любом документе найдётся хотя бы одна тема с ненулевой вероятностью.

Введём вспомогательный функционал

$$Q(\Phi, \Theta, \Phi', \Theta') = \sum_{d,w,t} n_{dw} p'_{tdw} \ln(\phi_{wt}\theta_{td}) + R(\Phi, \Theta), \qquad p'_{tdw} = \frac{\phi'_{wt}\theta'_{td}}{\sum\limits_t \phi'_{wt}\theta'_{td}}.$$

Это стандартный приём при доказательстве сходимости GEM алгоритма. Изменения Q на

итерациях, как будет показано в дальнейшем, являются нижней оценкой для изменений $L+R$. Аналогичный функционал вводился в статьях [3] и [17].

Теорема 2. *Пусть регуляризатор R является дифференцируемой функцией при $\phi_{wt}, \theta_{td} \in (0,1]$, сохраняющей нуль, корректной, ϵ-разреживающей и δ-регулярной. Также допустим, что $Q(\Phi^{k+1}, \Theta^{k+1}, \Phi^k, \Theta^k) \geq Q(\Phi^k, \Theta^k, \Phi^k, \Theta^k)$ начиная с некоторой итерации k. Тогда последовательность p_{tdw}^k сходится в смысле дивергенции Кульбака–Лейблера для любых d и w таких, что $n_{dw} > 0$:*

$$\mathrm{KL}\big(p_{tdw}^k \,\big\|\, p_{tdw}^{k+1}\big) \to 0 \text{ при } k \to \infty.$$

Доказательство. Поскольку регуляризатор сохраняет нуль, то, начиная с некоторой итерации, множество ячеек с нулевыми значениями в матрицах Φ и Θ стабилизируется и больше не будет изменяться. Это следует из того, что нулевое значение в ячейке не может стать на следующей итерации ненулевым, а множество всех ячеек конечно. Обозначим стабилизировавшееся множество ненулевых ячеек в матрицах Φ и Θ через Ω. Поскольку регуляризатор ϵ-разреживающий, значения Φ и Θ в позициях из Ω не могут быть менее ϵ. Но R — дифференцируемая функция при $\phi_{wt}, \theta_{td} \in [\epsilon, 1]$, следовательно, непрерывная и ограниченная.

Заметим, что Q можно переписать следующим образом:

$$Q(\Phi, \Theta, \Phi', \Theta') = L(\Phi, \Theta) + R(\Phi, \Theta) + \sum_{d,w,t} n_{dw} p'_{tdw} \ln p_{tdw}.$$

На M-шаге k-ой итерации были получены матрицы $(\Phi^{k+1}, \Theta^{k+1})$.

По условию теоремы, начиная с некоторой итерации выполнено

$$Q(\Phi^{k+1}, \Theta^{k+1}, \Phi^k, \Theta^k) \geq Q(\Phi^k, \Theta^k, \Phi^k, \Theta^k).$$

Подставим сюда вместо Q его выражение по определению:

$$\begin{aligned} L(\Phi^{k+1}, \Theta^{k+1}) + R(\Phi^{k+1}, \Theta^{k+1}) + \sum_{d,w,t} n_{dw} p_{tdw}^k \ln p_{tdw}^{k+1} \\ \geq L(\Phi^k, \Theta^k) + R(\Phi^k, \Theta^k) + \sum_{d,w,t} n_{dw} p_{tdw}^k \ln p_{tdw}^k, \end{aligned}$$

откуда следует

$$\Delta^k(L+R) \geq \sum_{d,w,t} n_{dw} p_{tdw}^k \ln \frac{p_{tdw}^k}{p_{tdw}^{k+1}} = \sum_{d,w} n_{dw} \mathrm{KL}\big(p_{dw}^k \,\big\|\, p_{dw}^{k+1}\big) \geq 0.$$

Равенство достигается только если на итерации не произошло никаких изменений, что означает, что процесс сошёлся в неподвижную точку. В обратном же случае $L+R$ строго увеличивается. Но это ограниченная функция, значит, $L(\Phi^k, \Theta^k) + R(\Phi^k, \Theta^k)$ сходится при $k \to \infty$. Более того $\mathrm{KL}\big(p_{tdw}^k \,\big\|\, p_{tdw}^{k+1}\big) \leq \Delta(L+R)^k \to 0$ при $n_{dw} > 0$, что завершает доказательство.

Следствие 1. *Если в дополнение к условиям Теоремы 2 регуляризатор R сильно регулярен, а r_{wt} и r_{td} непрерывны по параметрам модели, то*

$$|\phi_{wt}^k - \phi_{wt}^{k+1}| \to 0 \text{ и } |\theta_{td}^k - \theta_{td}^{k+1}| \to 0.$$

Доказательство. Согласно неравенству Пинскера [10], $\|A-B\|_1 \leq 2\sqrt{\mathrm{KL}(A\|B)}$. Поэтому сходимость по KL-дивергенции влечёт за собой сходимость по l_1 норме. Осталось заметить, что в условиях Теоремы 2 ϕ_{wt} и θ_{td} являются непрерывными функциями от p_{tdw}. Следовательно, сходимость вторых влечёт за собой сходимость первых.

Рассмотрим функцию $F(\Phi,\Theta) = L(\Phi,\Theta) + R(\Phi,\Theta)$, определённую для тех Φ и Θ, у которых множество нулевых позиций матриц совпадает с множеством ненулевых позиций Ω, стабилизировавшимся в ходе итераций.

Следствие 2. *В условиях Следствия 1 если процесс не сошёлся в неподвижную точку, то все предельные точки траектории (Φ^k,Θ^k) являются стационарными точками F. Если же множество стационарных точек F дискретно, то (Φ^k,Θ^k) сходится к некоторой стационарной точке F.*

Доказательство. В условиях Следствия 1 применение одной итерации EM-алгоритма к матрицам Φ и Θ является непрерывным преобразованием. Также в ходе доказательства теоремы было показано, что функция $F \equiv L + R$ строго возрастает на итерациях, если процесс не сошёлся в неподвижную точку. Остаётся заметить, что остальные условия Теоремы 1 тоже выполнены, если рассматривать все функции на области определения с ограничением на множество ненулевых позиций Ω.

Таким образом, итерационный процесс EM-алгоритма в ARTM разбиваются (в предположении увеличения Q) на два этапа: первый — выбор множества позиций Ω ненулевых ячеек в матрицах Φ и Θ, второй — окончательная оптимизация значений в этих ячейках. Первый этап можно рассматривать как дискретную оптимизацию структуры разреженности матриц Φ и Θ и подготовку их начальных приближений для второго этапа. Сходимость алгоритма происходит именно на втором этапе.

Таким образом, остаётся доказать монотонное увеличение функционала Q на втором этапе EM-алгоритма при фиксированном множестве Ω.

3. Изменение регуляризированного правдоподобия в EM-алгоритме

Важным условием сходимости алгоритма ARTM является неуменьшение значения Q на M-шаге. Далее будут приведены оценки изменения функционалов L, R и Q. Поскольку мы рассматриваем второй этап итерационного процесса, когда множество нулевых позиций в матрицах Φ и Θ не изменяется, положительную срезку в формулах можно опустить.

Введём функционал $\bar{Q}(\Phi,\Theta,\Phi',\Theta') = \sum\limits_{d,w,t} n_{dw} p'_{tdw} \ln(\phi_{wt}\theta_{td})$. Тогда $Q = \bar{Q} + R$.

Провести анализ суммарного изменения функционала Q на M-шаге напрямую затруднительно. Поэтому предлагается разложить это преобразование на два этапа. Первый этап — максимизация $\bar{Q}$:

$$\begin{cases} \phi_{wt} = \underset{w\in W}{\mathrm{norm}}(n_{wt}), \\ \theta_{td} = \underset{t\in T}{\mathrm{norm}}(n_{td}). \end{cases}$$

Второй этап (назовём его регуляризационным преобразованием) — максимизация R:

$$\begin{cases} \phi_{wt} = \underset{w\in W}{\mathrm{norm}}\left(n_{wt} + r_{wt}\right), \\ \theta_{td} = \underset{t\in T}{\mathrm{norm}}\left(n_{td} + r_{td}\right) \end{cases} \tag{3.1}$$

Таким образом, изменения функционалов будут оцениваться отдельно на каждом этапе. На первом происходит переход в точку $(n_{wt}/n_t, n_{td}/n_d)$, которая является точкой максимума функционала $\bar{Q}$, а на втором проводится максимизация R.

Введём ещё один функционал и обозначения для его частных производных:

$$\bar{R}\big((m_{wt}),(m_{td})\big) = R\bigg(\frac{m_{wt}}{\sum\limits_w m_{wt}}, \frac{m_{td}}{\sum\limits_t m_{td}}\bigg) = R\left(\frac{m_{wt}}{m_t}, \frac{m_{td}}{m_d}\right);$$

$$g_{wt} \equiv \frac{\partial \bar{R}}{\partial m_{wt}}, \quad g_{td} \equiv \frac{\partial \bar{R}}{\partial m_{td}}, \quad \phi_{wt} = \frac{m_{wt}}{\sum\limits_w m_{wt}}, \quad \theta_{td} = \frac{m_{td}}{\sum\limits_t m_{td}}.$$

Таким образом, функционал $\bar{R}$, определён на паре произвольных неотрицательных матриц размера $|W| \times |T|$ и $|T| \times |D|$. Он нормирует эти матрицы и применяет к ним регуляризатор R. Отметим, что при регуляризационном преобразовании $\bar{R}(n_{wt}, n_{td}) = R\big(n_{wt}/n_t, n_{td}/n_d\big)$.

Утверждение 1. *Для g_{wt} и g_{td} выполнено:*

$$g_{wt} = \frac{1}{m_t} \sum_{u \in W} \left(\frac{\partial R}{\partial \phi_{wt}} - \frac{\partial R}{\partial \phi_{ut}}\right) \phi_{ut},$$

$$g_{td} = \frac{1}{m_d} \sum_{s \in T} \left(\frac{\partial R}{\partial \theta_{td}} - \frac{\partial R}{\partial \theta_{sd}}\right) \theta_{sd}.$$

Доказательство. В силу нормировки $\phi_{wt} = \frac{m_{wt}}{\sum_w m_{wt}}$,

$$\frac{\partial \phi_{ut}}{\partial m_{wt}} = \frac{\partial \frac{m_{ut}}{\sum\limits_v m_{vt}}}{\partial m_{wt}} = \frac{\frac{\partial m_{ut}}{\partial m_{wt}}}{\sum\limits_v m_{vt}} - \frac{m_{ut}}{(\sum\limits_v m_{vt})^2} = \frac{[u=w]}{m_t} - \frac{\phi_{ut}}{m_t} = \frac{1}{m_t}\big([u=w] - \phi_{ut}\big).$$

Следовательно,

$$\frac{\partial \bar{R}}{\partial m_{wt}} = \sum_u \frac{\partial R}{\partial \phi_{ut}} \frac{\partial \phi_{ut}}{\partial m_{wt}} = \frac{1}{m_t}\left(\frac{\partial R}{\partial \phi_{wt}} - \sum_u \frac{\partial R}{\partial \phi_{ut}} \phi_{ut}\right) = \frac{1}{m_t} \sum_u \left(\frac{\partial R}{\partial \phi_{wt}} - \frac{\partial R}{\partial \phi_{ut}}\right) \phi_{ut}.$$

Формула для g_{dt} доказывается аналогично.

Теперь докажем основную теорему.

Теорема 3. *Пусть величины r_{wt} и r_{td} на M-шаге рассчитываются в точках*

$$\frac{n_{wt}}{\sum\limits_w n_{wt}} \text{ и } \frac{n_{td}}{\sum\limits_t n_{td}},$$

тогда в ходе регуляризационного преобразования (3.1) *без занулений элементов матриц угол между вектором изменений и градиентом R острый, если градиент ненулевой.*

Доказательство. Докажем утверждение для Δn_{wt}, для Δn_{td} доказательство будет аналогично. При регуляризационном преобразовании без занулений $\Delta n_{wt} = \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}$, поэтому с учётом Утверждения 1 получаем:

$$\langle \Delta n, \nabla \bar{R}(n_{wt}, n_{td}) \rangle = \sum_{w,t,u} \frac{1}{n_t} \left(\frac{\partial R}{\partial \phi_{wt}} - \frac{\partial R}{\partial \phi_{ut}}\right) \frac{\partial R}{\partial \phi_{wt}} \phi_{wt} \phi_{ut}.$$

В силу симметрии суммы выполнено:

$$\sum_{w,t,u}\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{ut}}\right)\frac{\partial R}{\partial\phi_{wt}}\phi_{wt}\phi_{ut}=\sum_{w,t,u}\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{ut}}-\frac{\partial R}{\partial\phi_{wt}}\right)\frac{\partial R}{\partial\phi_{ut}}\phi_{wt}\phi_{ut}$$

$$=\sum_{w,t,u}\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{ut}}\right)\left(-\frac{\partial R}{\partial\phi_{ut}}\right)\phi_{wt}\phi_{ut}$$

$$=\frac{1}{2}\left(\sum_{w,t,u}\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{ut}}\right)\frac{\partial R}{\partial\phi_{wt}}\phi_{wt}\phi_{ut}+\sum_{w,t,u}\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{ut}}\right)\left(-\frac{\partial R}{\partial\phi_{ut}}\right)\phi_{wt}\phi_{ut}\right)$$

$$=\frac{1}{2}\sum_{t,w,u}\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{ut}}\right)^2\phi_{wt}\phi_{ut}=\sum_{t,w<u}\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{ut}}\right)^2\phi_{wt}\phi_{ut}\geq 0.$$

Пусть здесь достигается равенство, тогда $\frac{\partial R}{\partial\phi_{wt}}=\frac{\partial R}{\partial\phi_{ut}}$ для всех u и w. Тогда

$$\frac{\partial\bar{R}}{\partial n_{wt}}=\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\sum_u\frac{\partial R}{\partial\phi_{ut}}\phi_{ut}\right)=\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\sum_u\frac{\partial R}{\partial\phi_{wt}}\phi_{ut}\right)$$

$$=\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{wt}}\sum_u\phi_{ut}\right)=\frac{1}{n_t}\left(\frac{\partial R}{\partial\phi_{wt}}-\frac{\partial R}{\partial\phi_{wt}}\right)=0.$$

Значит, градиент нулевой. Получили противоречие. Поэтому неравенство строгое и угол острый, что и требовалось доказать.

Ранее было показано (Теорема 2), что при определённых ограничениях на регуляризатор занулений ячеек в матрицах Φ и Θ не будет, начиная с некоторой итерации. Таким образом, если коэффициенты регуляризации не слишком большие, то изменение n_{wt} и n_{td} будет незначительно. Поэтому при регуляризационном преобразовании будет происходить увеличение R в силу локального изменения вдоль градиента.

Теперь нужно объединить результаты двух этапов. В ходе первого этапа происходит переход в точку максимума $\bar{Q}$, значит, градиент $\bar{Q}$ в этой точке нулевой. Это означает, что в ней градиент $\bar{Q}+R$ сонаправлен с градиентом R, откуда следует, что на этапе регуляризационного преобразования происходит неуменьшение $\bar{Q}+R$.

Остаётся понять, как изменяется этот функционал на первом этапе. Есть риск, что при максимизации $\bar{Q}$ значение Q может уменьшиться, поэтому при реализации алгоритма необходимо дополнительно проверять, что значение Q увеличилось на итерации и использовать новое значение Φ и Θ только если увеличение произошло. Эта проверка строго гарантирует неуменьшение Q на итерациях.

4. Модификация М-шага

Обычно в реализациях ЕМ-алгоритма для ARTM [1; 4; 12] регуляризационные поправки r_{wt} и r_{td} рассчитываются в точке (Φ^k,Θ^k). В этом случае нет теоретических гарантий на увеличение Q на этапе регуляризационного преобразования. Поэтому алгоритм может сойтись в неподвижную точку отображения, а не в стационарную точку функционала $L+R$, из-за чего значение $L+R$ окажется субоптимальным.

Теорема 3 утверждает, что если рассчитывать r_{wt} и r_{td} в точке $\big((n^k_{wt}/n^k_t),(n^k_{td}/n^k_d)\big)$, то есть на основе величин, подсчитанных на М-шаге k-й итерации, то будут выполнены теоретические гарантии оптимальности.

Таким образом, обычные формулы M-шага для регуляризационных поправок

$$r_{wt}^k = \phi_{wt}^{k-1}\frac{\partial R}{\partial \phi_{wt}}(\Phi_{wt}^{k-1}, \Theta_{td}^{k-1}); \qquad r_{td}^k = \theta_{td}^{k-1}\frac{\partial R}{\partial \theta_{td}}(\Phi_{wt}^{k-1}, \Theta_{td}^{k-1}); \tag{4.2}$$

заменяются на модифицированные согласно Теореме 3:

$$r_{wt}^k = \frac{n_{wt}^k}{n_t^k}\frac{\partial R}{\partial \phi_{wt}}\left(\frac{n_{wt}^k}{n_t^k}, \frac{n_{td}^k}{n_d^k}\right); \qquad r_{td}^k = \frac{n_{td}^k}{n_d^k}\frac{\partial R}{\partial \theta_{wt}}\left(\frac{n_{wt}^k}{n_t^k}, \frac{n_{td}^k}{n_d^k}\right). \tag{4.3}$$

В следующем разделе будет проведено сравнение этих двух версий EM-алгоритма на реальной текстовой коллекции.

5. Эксперимент

Согласно Теореме 3, если рассчитывать регуляризационные поправки r_{wt} и r_{td} не по матрицам Φ_{wt} и Θ_{td} с предыдущей итерации, а по матрицам (n_{wt}) и (n_{td}), то значение оптимизируемого функционала будет гарантированно увеличиваться на втором этапе итерационного процесса. Ожидается, что это ускорит оптимизацию, позволяя за то же число итераций получать лучшие значения максимизируемого критерия.

Для экспериментальной проверки этого утверждение мы использовали лемматизированную коллекцию новостных сообщений на английском языке «20 NewsGroups» [7]. Тематическая модель строилась EM-алгоритмом для ARTM, описанным в [13], с использованием регуляризатора декоррелирования [8]:

$$R(\Phi) = -\frac{\tau}{|T|(|T|-1)}\sum_{t\neq s}\sum_{w\in W}\phi_{wt}\phi_{ws}.$$

Данный регуляризатор был выбран как один из наиболее часто используемых. Его максимизация способствует увеличению попарной различности тем как столбцов матрицы Φ, улучшает интерпретируемость тем и способствует выделению фоновых тем с общей лексикой языка. При этом регуляризатор декоррелирования не имеет аналитического решения для задачи максимизации функционала Q на M-шаге.

В эксперименте мы проверяли, как на итерациях алгоритма изменяется значение оптимизируемого функционала $L(\Phi, \Theta) + R(\Phi)$. Значения τ перебирались в таком интервале, чтобы абсолютная величина R была соизмерима с абсолютным значением L и регуляризатор оказывал заметное влияние на модель в процессе оптимизации. Сравнивались две версии M-шага: стандартная (4.2) и модифицированная (4.3).

На Рис. 1 видно, что при стандартных формулах M-шага на первых итерациях происходит уменьшение функционала $L+R$, причём с ростом τ количество таких итераций растёт. В то же время для модифицированного шага только одна итерация происходит с уменьшением $L+R$, далее наблюдается рост значений. Как и предполагалось, это позволяет получить заметно лучшие значения $L+R$ в точке, к которой сходится алгоритм. Их сравнение приводится в таблице, Рис.2. Также заметим, что чем больше τ, то есть чем сильнее воздействие регуляризатора на модель, тем существеннее предложенная модификация улучшает полученное решение.

6. Заключение

Данная работа закрывает проблему обоснования сходимости EM-алгоритма в ARTM при произвольном гладком критерии регуляризации. Полученные ограничения на регуляризатор

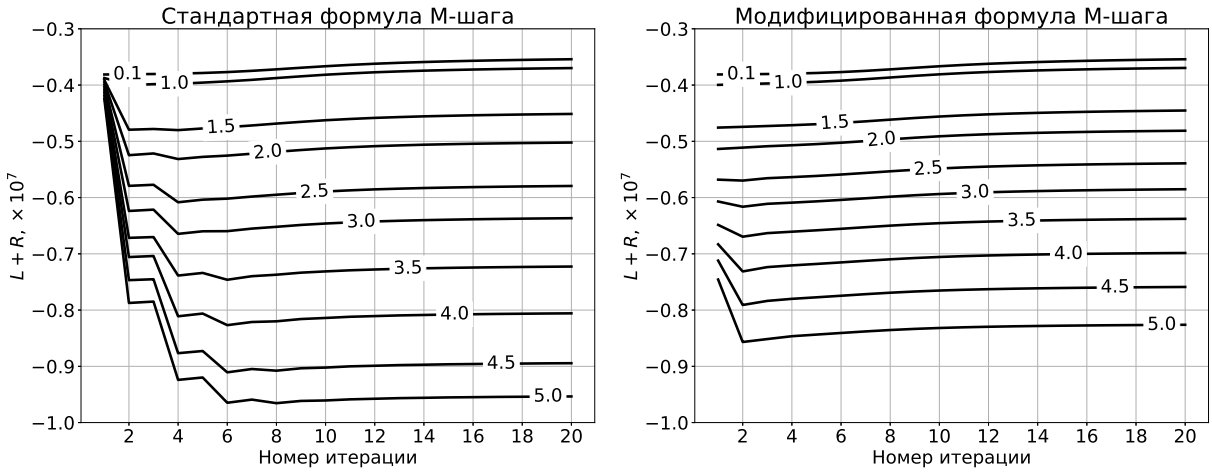

Рис. 1. Изменение функционала $L+R$ на итерациях, $|T|=30$, при различных значениях коэффициента регуляризации τ (графики подписаны значениями τ, уменьшенными в 10^8 раз).

τ	$L+R$ стандарт	$L+R$ модификация	Увеличение $L+R$, %
10^7	-3536050	-3536340	-0.01
10^8	-3693905	-3691338	0.07
$1.5 \cdot 10^8$	-4509247	-4448501	1.35
$2.0 \cdot 10^8$	-5018335	-4808217	4.19
$2.5 \cdot 10^8$	-5790283	-5388187	6.94
$3.0 \cdot 10^8$	-6363392	-5848354	8.09
$3.5 \cdot 10^8$	-7223361	-6374974	11.75
$4.0 \cdot 10^8$	-8055262	-6982549	13.32
$4.5 \cdot 10^8$	-8941616	-7586618	15.15
$5.0 \cdot 10^8$	-9532948	-8259205	13.36

Рис. 2. Итоговые значения функционала $L+R$ по окончании итераций.

не являются обременительными, легко проверяются и легко обеспечиваются программной реализацией. Весьма неожиданным оказался тот факт, что итерационный процесс, обычно используемый для построения регуляризованных моделей, в общем случае не гарантирует сходимости к стационарной точке. Модификация EM-алгоритма, исправляющая этот недостаток, не требует дополнительных затрат времени или памяти. Она сводится к тому, чтобы при вычислении регуляризационных поправок вместо текущих значений условных вероятностей ϕ_{wt}, θ_{td} подставлять их нерегуляризованные частотные оценки — ровно те, которые вычисляются в модели PLSA.

СПИСОК ЛИТЕРАТУРЫ

1. **Apishev M. A., Vorontsov K. V.** Learning topic models with arbitrary loss // Proceeding of the 26th Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. 2020. Pp. 30–37.

2. **Blei D. M., Ng A. Y., Jordan M. I.** Latent Dirichlet allocation // the Journal of machine Learning research. 2003. Vol. 3. Pp. 993–1022.
3. **Dempster A. P., Laird N. M., Rubin D. B.** Maximum likelihood from incomplete data via the EM algorithm // Journal of the royal statistical society. Series B (methodological). 1977. Pp. 1–38.
4. **Frei O. I., Apishev M. A.** Parallel non-blocking deterministic algorithm for online topic modeling // AIST'2016, Analysis of Images, Social networks and Texts. Vol. 661. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2016. Pp. 132–144.
5. **Hofmann T.** Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval / ACM. 1999. Pp. 50–57.
6. **Kochedykov D. A., Apishev M. A., Golitsyn L. V., Vorontsov K. V.** Fast and modular regularized topic modelling // Proceeding of The 21st Conference of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6–10, 2017. IEEE, 2017. Pp. 182–193.
7. **Lang K.** 20 newsgroups. 2008. Data retrieved from the dataset's official website, `http://qwone.com/~jason/20Newsgroups/`.
8. **Tan Y., Ou Z.** Topic-weak-correlated latent Dirichlet allocation // 7th International Symposium Chinese Spoken Language Processing (ISCSLP). 2010. Pp. 224–228.
9. **Tikhonov A. N., Arsenin V. Y.** Solution of ill-posed problems. W. H. Winston, Washington, DC, 1977.
10. **Topsøe F.** Some inequalities for information divergence and related measures of discrimination // Information Theory, IEEE Transactions on. 2000. Vol. 46, no. 4. Pp. 1602–1609.
11. **Vorontsov K. V.** Additive regularization for topic models of text collections // Doklady Mathematics. 2014. Vol. 89, no. 3. Pp. 301–304.
12. **Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A.** BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST'2015, Analysis of Images, Social networks and Texts. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2015. Pp. 370–384.
13. **Vorontsov K. V., Potapenko A. A.** Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, Analysis of Images, Social networks and Texts. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. Vol. 436. Pp. 29–46.
14. **Vorontsov K. V., Potapenko A. A.** Additive regularization of topic models // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications. 2015. Vol. 101, no. 1. Pp. 303–323.
15. **Vorontsov K. V., Potapenko A. A., Plavin A. V.** Additive regularization of topic models for topic selection and sparse factorization // The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK. Springer International Publishing Switzerland 2015, 2015. Pp. 193–202.
16. **Wallach H. M., Mimno D. M., McCallum A.** Rethinking LDA: Why priors matter // Advances in neural information processing systems. 2009. Pp. 1973–1981.
17. **Wu C. J.** On the convergence properties of the EM algorithm // The Annals of statistics. 1983. Pp. 95–103.

Поступила 20.07.2020

Ирхин Илья Александрович,
аспирант, Московский физико-технический институт, г.Москва,
e-mail: ilirhin@gmail.com

Воронцов Константин Вячеславович,
д.ф.-м.н., профессор РАН, зав. лаб. машинного интеллекта МФТИ,
Московский физико-технический институт, г.Москва,
e-mail: k.v.vorontsov@phystech.edu