Московский физико-технический институт (ГУ)

На правах рукописи

Ирхин Илья Александрович

Единственность матричного разложения и сходимость регуляризованных алгоритмов в вероятностном тематическом моделировании

Специальность 05.13.17 —
«Теоретические основы информатики»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук
Воронцов Константин Вячеславович

Москва — 2020

Оглавление

Стр.

Стр.

Стр.

Введение

Актуальность исследования. Развитие современных технологий сбора и хранения информации привело к заметному увеличению объёмов данных, в частности текстовых документов. Во многих прикладных областях возникает потребность в обработке и анализе накопленных текстовых коллекций. Одним из популярных в настоящее время направлений обработки естественного языка (Natural Language Processing, NLP) является тематическое моделирование. Тематическая модель описывает зависимость текстовых документов и содержащихся в них термов через наборы кластеров термов — тем. В роли термов обычно рассматриваются слова или их нормальные формы, но также иногда используются словосочетания или термины. Конкретная форма термов зависит то того, какие виды предварительной обработки текста были применены к коллекции. Тематическая модель для текстовой коллекции относит каждый документ к некоторым темам и для каждой темы определяет какие термы её образуют. Выявление подобных тематики текста можно рассматривать как шаг в направлении понимания естественного языка (Natural Language Understanding, NLU). В частности, тематическое моделирование предоставляет вариант решения проблемы синонимии и полисемии слов. Синонимы объединяются в одну тему, поскольку обычно употребляются в схожих контекстах. В то же время слова с несколькими значениями и омонимы попадают сразу в несколько тем, позволяя отличить разные по смыслу употребления друг от друга.

Тематическое моделирование может быть использовано для получения интерпретируемых векторных представлений слов, демонстрирующих сравнимое качество с векторными представлениями модели SGNS (Skip-Gram Negative Sampling) [1] на задачах сравнения семантически близких слов [2]. Но применение тематического моделирования не ограничивается только областью анализа текстов. Данный подход применяется и в других областях, например, в анализе аудио [3], анализе изображений и видео [4—6], биоинформатике [7; 8]. Также тематические модели используются в задачах информационного поиска [9—12] и рекомендаций [13—15].

Общий подход для решения задачи тематического моделирования — построение вероятностной тематическая модели (Probabilistic Topic Model, PTM). Согласно этому подходу документы описываются некоторым дискретным

распределением вероятностей на множестве тем, а темы — 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.

Классическим методом построения PTM является предложенный в 1999 году вероятностный латентный семантический анализ (Probabilistic Latent Semantic Analysis, PLSA [16]). Этот подход задаёт вероятностную модель порождения термов в документах и строит разбиение термов и документов на темы, исходя из принципа максимизации правдоподобия. В 2003 году была предложена классическая модель латентного размещения Дирихле (Latent Dirichlet Allocation, LDA [17]), которая была более устойчива и более точно учитывала данные о редких термах.

Важным преимуществом модели LDA является возможность расширять вероятностную модель дополнительными параметрам. Это сыграло важную роль в популярности подхода LDA [18]. Именно на основе этого подхода, были предложены вероятностные модели, учитывающие связи между документами [19—21] или метаданные о документах [22]. Также есть модели, которые учитывают время появления документа и его язык [23] или порядок слов в документе [24; 25], что изначально было несвойственно подходу LDA.

Традиционный способ построения новых тематических моделей описан в [18]. Рекомендации включают в себя: введение новой вероятностной модели коллекции документов, которая не должна быть слишком вычислительно сложной, оставаясь при этом реалистичной; нахождение нового алгоритма оценки апостериорного распределения параметров; реализацию этого алгоритма; валидацию результатов. Трудностями использования подобного подхода являются необходимость проделывать данные действия заново для каждой новой модели, а также сложность и, иногда, невозможность построения тематических моделей, удовлетворяющих нескольким различным требований одновременно.

Теория аддитивной регуляризации тематических моделей (Additive Regularization of Topic Models, ARTM) [26] решает эти проблемы, отказываясь от использования байесовского вывода [27]. В ARTM любые требования к модели формализуются через оптимизационные критерии — регуляризаторы.

Если требований несколько, то в постановку оптимизационной задачи вводится взвешенная сумма регуляризаторов [26; 28]. Байесовские тематические модели, как правило, удаётся переформулировать в терминах регуляризации, при этом существенно сокращается объём необходимых математических выкладок [29]. Для оценивания параметров модели с произвольным набором регуляризаторов используется один и тот же итерационный процесс, называемый регуляризованным EM-алгоритмом. Этот алгоритм даёт возможность добавлять и заменять регуляризаторы не только на уровне постановки задачи, но и на уровне алгоритма и его программного кода. Это приводит к модульной технологии тематического моделирования, которая реализована в проектах с открытым кодом `BigARTM` [30; 31] и `TopicNet` [32].

Дополнительным обоснованием использования регуляризаторов является некорректность по Адамару [33] поставленной оптимизационной задачи максимизации правдоподобия. Согласно теории регуляризации А. Н. Тихонова [34], добавление регуляризатора доопределяет решение задачи и делает его устойчивым.

До сих пор в теории ARTM оставались открытыми вопросы о сходимости регуляризованного EM-алгоритма и о влиянии регуляризаторов на сходимость. В литературе хорошо изучены свойства Generalized Expectation Maximization алгоритма (GEM [35]), для которого известны достаточные условия сходимости [36]. В данной работе показывается, что итерации регуляризованного EM-алгоритма ARTM возможно интерпретировать как итерации GEM-алгоритма, за счёт чего возможно получить достаточные условия сходимости.

Для сходящегося итерационного процесса ARTM ставится вопрос о свойствах точки, к которой сошёлся алгоритм, например, открытым является вопрос о единственности полученного решения. С формальной точки зрения, в алгоритме ARTM для матрицы частот слов в документах строится стохастическое матричное разложение, ранг которого равен числу тем. Поскольку точного разложения требуемого ранга, как правило, не существует, то строится приближенное разложение, которое является локальным экстремумом оптимизируемого функционала. Таким образом, неединственность решения задачи тематического моделирования может возникать как из-за неоднозначности выбора этого приближения, так и из-за неединственности точного разложения приближения. До сих пор остаётся открытым вопрос о влиянии этих факторов на неединственность решения задачи тематического моделирования.

Проблема единственности стохастического матричного разложения исследовалась в работах [37—39]. В этих работах представлены либо достаточные, либо необходимые условия единственности разложения. Недостатками предложенных условий применительно к тематическому моделированию являются их громоздкость и сложность проверки выполнения на практике.

В данной работе исследуются теоретические свойства регуляризованного EM-алгоритма ARTM. Особое внимание уделяются вопросам сходимости данного алгоритма и единственности стохастического матричного разложения в точке сходимости, поскольку они являются открытыми и представляют отдельный интерес. Также в рамках исследования производится поиск возможных модификаций алгоритма, которые за счёт теоретических гарантий будут улучшать качество получаемых тематических моделей.

Целью данной работы является получение достаточных условий сходимости алгоритма аддитивной регуляризации тематических моделей и достаточных условий для единственности стохастического матричного разложения в точке сходимости, которые могут быть проверены на реальных текстовых коллекциях, а также поиск модификаций исходного алгоритма, улучшающих сходимость и повышающих метрики качества тематических моделей.

Методология и методы исследования. В работе использованы подходы и методы численной оптимизации, вычислительной линейной алгебры, теории матричных разложений, машинного обучения. Для доказательства сходимости алгоритма ARTM использовались известные фундаментальные результаты о сходимости GEM-алгоритмов. Для доказательства единственности стохастического матричного разложения был использован подход геометрической интерпретации стохастического матричного разложения. В качестве реализации алгоритма ARTM использовались собственная реализация на языке Python[1] а также библиотеки с открытым кодом BigARTM [31] и TopicNet [32]. Для экспериментов в качестве текстовых коллекций использовались открытые публичные данные.

Научная новизна:

1. Впервые были получены достаточные условия сходимости алгоритма аддитивной регуляризации тематических моделей ARTM.
2. Были получены достаточные условия единственности стохастического матричного разложения в задачах тематического моделирования.

[1] github.com/ilirhin/python_artm/tree/master/

3. Были сформулированы причины нединственности решения для задач тематического моделирования.
4. Был разработан новый подход к стохастическому матричному разложению в тематическом моделировании, в котором одна из матриц находится в функциональной зависимости от другой.

Теоретическая значимость В работе впервые предложен подход с интерпретацией ARTM как GEM-алгоритма, в результате чего были получены достаточные условия сходимости данного алгоритма. Также были получены достаточные условия на единственность стохастического матричного разложения. В результате были сформулированы причины неединственности решения в тематическом моделировании.

Практическая значимость Разработана реализация алгоритма ARTM, с помощью которой теоретические положения диссертационной работы были подтверждены на реальных текстовых коллекциях. Предложенные в работе в работе алгоритмы реализованы в библиотеке с открытым кодом TopicNet. Модификации EM-алгоритма ARTM, полученные на основе теоретических результатов, значительно увеличивают основные метрики качества тематических моделей.

Основные положения, выносимые на защиту:

1. Теорема о достаточных условиях сходимости алгоритма ARTM.
2. Теорема о достаточных условиях единственности стохастического матричного разложения.
3. Модификация алгоритма ARTM, ускоряющая сходимость итерационного процесса.
4. Метод разреживания тематической модели, не увеличивающий перплексию получаемой модели.

Достоверность Достоверность результатов обеспечивается доказательствами теорем и описаниями проведённых экспериментов, допускающими их воспроизводимость, а также наличием репозитория Github с исходным кодом всех экспериментов.

Апробация работы. Основные результаты работы докладывались на:

1. 5th International Symposium, Conformal and Probabilistic Prediction with Applications, 2016
2. Научный семинар Школы Анализа Данных, 2016.
3. Научный семинар лаборатории искусственного интеллекта, 2018.

4. Научный семинар Федерального исследовательского центра «Информатика и Управление» Российской Академии Наук, 2020.

Личный вклад. Личный вклад диссертанта в работы, выполненные с соавторами, заключается в следующем:

- В работе [40] предложена идея применения подхода тематического моделирования, предложены метрики качества, соответствующие решению прикладной задачи.
- В работе [41] предложены достаточные условия сходимости, доказаны все утверждения и теоремы, реализованы и проведены все эксперименты.
- В работе [42] предложена и доказана основная лемма, реализованы и проведены все эксперименты.
- В работе [43] предложена новая постановка оптимизационной задачи, выполнен вывод итераций алгоритма ARTM, реализована и проведена часть экспериментов, не связанная с библиотекой TopicNet.

Основные результаты по теме диссертации изложены в 3 печатных изданиях, 1 из которых изданы в журналах, рекомендованных ВАК, 2 — в периодических научных журналах, индексируемых Scopus.

Объем и структура работы. Диссертация состоит из введения, 5 глав, заключения и 0 приложений. Полный объём диссертации составляет 105 страниц, включая 14 рисунков и 7 таблиц. Список литературы содержит 76 наименований.

Глава 1. Аддитивная регуляризация тематических моделей

В этой главе будут поставлены оптимизационные задачи вероятностного тематического моделирования и аддитивной регуляризации тематических моделей. Также будут введены основные общие обозначения, которые будут использоваться в данной работе.

1.1 Постановка задачи тематического моделирования

Пусть D — множество (коллекция) текстовых документов, W — множество (словарь) всех употребляемых в них терминов (слов или словосочетаний). Каждый документ $d \in D$ представляет собой последовательность n_d терминов $(w_1, ..., w_{n_d})$ из словаря W. Термин может повторяться в документе много раз. Пусть существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$, которая не известна. Формально, тема определяется как дискретное (мультиномиальное) вероятностное распределение в пространстве слов заданного словаря W.

Вводится дискретное вероятностное пространство $D \times W \times T$. Тогда коллекция документов может быть рассмотрена как множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$. При этом документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, темы $t \in T$ являются латентными (скрытой) переменными. Через n_{dw} обозначается число вхождений терма w в документ d.

Пусть $\varphi_{wt} = p(w\,|\,t)$ — неизвестное распределение термов в темах, $\theta_{td} = p(t\,|\,d)$ — неизвестное распределения тем в документах. Задача вероятностного тематического моделирования заключается в том, чтобы найти параметры модели по эмпирическим данным n_{dw}. Эта задача решается с помощью метода максимизации логарифма правдоподобия.

Сначала принимается гипотеза условной независимости, утверждающая, что $p(w|d,t) = p(w|t)$, и по формуле полной вероятности определяется вероят-

ность появления слова w в документе d:

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

Тогда логарифм правдоподобия коллекции D определяется как

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td}.$$

Таким образом, для нахождения параметров модели решается задача максимизации логарифма правдоподобия

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

при ограничениях неотрицательности и нормировки:

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

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

Данная оптимизационная задача решается при помощи EM-алгоритма [16], 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}:

$$p_{tdw} \equiv p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s\in T} \varphi_{ws}\theta_{sd}} \tag{1.2}$$

На M-шаге по условным вероятностям тем p_{tdw} для каждого терма в каждом документе вычисляются новые приближения параметров φ_{wt} и θ_{td}:

$$\varphi_{wt} = \frac{\sum\limits_{d\in D} n_{dw}p_{tdw}}{\sum\limits_{d\in D}\sum\limits_{w\in W} n_{dw}p_{tdw}}, \theta_{td} = \frac{\sum\limits_{w\in W} n_{dw}p_{tdw}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dw}p_{tdw}} \tag{1.3}$$

1.2 Регуляризация тематических моделей

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 предлагается максимизировать взвешенную сумму всех регуляризаторов $R(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta)$ совместно с основным критерием правдоподобия:

$$L(\Phi,\Theta)+R(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \log \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}, \quad (1.4)$$

при тех же ограничениях неотрицательности и нормировки что и в (1.1).

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

$$R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W} (\beta_w - 1) \ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1) \ln \theta_{td} \quad (1.5)$$

с гиперпараметрами β_w, α_t.

Применение теоремы Каруша–Куна–Таккера позволяет выписать систему уравнений для стационарных точек оптимизационной задачи (1.4).

Теорема 1. *Решение* Φ,Θ *задачи* (1.4) *при ограничениях неотрицательности и нормировки удовлетворяет следующей системе уравнений относительно переменных* φ_{wt}, θ_{td} *и вспомогательных переменных* p_{tdw}*:*

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big);$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

Решение данной системы методом простых итераций приводит к EM-подобному алгоритму, E-шаг которого аналогичен (1.2), а M-шаг изменяется на

$$
\begin{aligned}
n_{dwt} &= n_{dw} p_{tdw}, \\
n_{wt} &= \sum_{d \in D} n_{dwt}, & n_{td} &= \sum_{w \in d} n_{dwt}, \\
n_t &= \sum_{w \in W} n_{wt}, & n_d &= \sum_{t \in T} n_{td}, \\
r_{wt} &= \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}, & r_{td} &= \theta_{td} \frac{\partial R}{\partial \theta_{td}}, \\
\varphi_{wt} &= \underset{w \in W}{\operatorname{norm}} \left(n_{wt} + r_{wt} \right), & \theta_{td} &= \underset{t \in T}{\operatorname{norm}} \left(n_{td} + r_{td} \right),
\end{aligned}
\tag{1.6}
$$

где $\underset{i \in I}{\operatorname{norm}}(x_i) = \frac{(x_i)_+}{\sum_{j \in I}(x_j)_+}$ — 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:

n_{dwt} — числа вхождений терма w в документ d, связанных с темой t;

n_{td} — числа всех термов в документе d, связанных с темой t;

n_{wt} — числа раз, когда терм w был связан с темой t, во всей коллекции;

n_t — числа термов, связанных с темой t, во всей коллекции;

n_d совпадает с длиной документа d.

Чаще всего используются следующие регуляризаторы:

1. $R(\Phi, \Theta) = \alpha \sum_{w,t} \ln \varphi_{wt}$ — регуляризатор сглаживания.
2. $R(\Phi, \Theta) = -\alpha \sum_{w,t} \ln \varphi_{wt}$ — регуляризатор разреживания.
3. $R(\Phi, \Theta) = -\tau \sum_{w \neq u, t} \varphi_{wt} \varphi_{ut}$ — регуляризатор декоррелирования.
4. $R(\Phi, \Theta) = \sum_{w \neq u, t} C_{uw} \left(\varphi_{wt} - \varphi_{ut} \right)^2$ — регуляризатор когерентности.
5. $R(\Phi, \Theta) = \sum_{s \neq t, d} C_{st} \left(\theta_{td} - \theta_{sd} \right)^2$ — регуляризатор связей документов (лапласиан графа документов).

Более подробное описание данных регуляризаторов, а также другие регуляризаторы можно найти в работах [26; 44; 45]

Для комбинирования регуляризаторов при решении задачи в APTM необходимо продумывать стратегию регуляризации:

1. Какие регуляризаторы необходимы в данной задаче.

2. Какие регуляризаторы должны работать одновременно, какие друг за другом или попеременно, делая необходимую подготовительную работу.
3. Как менять коэффициент регуляризации каждого регуляризатора в ходе итераций: по каким условиям включать, усиливать, ослаблять и отключать каждый регуляризатор.

1.3 Обобщение для произвольных функций потерь

Для оптимизационной задачи (1.4) рассматривается следующее обобщение. Вводится функция потерь $\ell\big(p(w|d)\big)$ и ставится оптимизационная задача:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ell\Big(\sum_{t\in T}\varphi_{wt}\theta_{td}\Big) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}, \tag{1.7}$$

где $\ell(z)$ — произвольная гладкая неубывающая функция.

По аналогии с Теоремой 1 верна следующая теорема:

Теорема 2. *Решение* Φ,Θ *задачи* (1.7) *при ограничениях неотрицательности и нормировки удовлетворяет следующей системе уравнений относительно переменных* φ_{wt}, θ_{td} *и вспомогательных переменных* p_{tdw}*:*

$$\begin{aligned}
p_{tdw} &= \varphi_{wt}\theta_{td}\ell'\big(p(w|d)\big);\\
\varphi_{wt} &= \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}\Big);\\
\theta_{td} &= \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Big).
\end{aligned}$$

Решение данной системы уравнений методом простых итераций отличается от классической только формулой E-шага.

При $\ell(z) = \ln z$ на E-шаге оптимизационная задача (1.7) совпадает с оптимизационной задачей (1.4) и E-шаг алгоритма совпадает с E-шагом ARTM и PLSA (1.2).

При $\ell(z) = z$ вместо правдоподобия максимизируется суммарная близость модельных распределений вероятности термов в документах $p(w|d)$ и эмпирических распределений $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$, выраженная через скалярные произведения:

$$\sum_{d \in D} n_d \big\langle \hat{p}(w|d), p(w|d) \big\rangle + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}.$$

При этом $p_{tdw} = \varphi_{wt}\theta_{td}$, то есть из классической формулы E-шага уходит нормировочный множитель в знаменателе. Этот случай функции потерь называется быстрым E-шагом.

Быстрый E-шаг даёт существенное ускорение EM-алгоритма, поскольку в классическом варианте вычисление нормировочного множителя

$$p(w\,|\,d) = \sum_{t \in T} \varphi_{wt}\theta_{td} = \langle \varphi_w, \theta_d \rangle$$

занимает больше всего времени при обработке каждого терма в каждом документе.

1.4 Алгоритм ARTM в матричной форме

При экспериментах с алгоритмом ARTM, особенно при тестировании его модификаций, важна производительность алгоритма, а с другой стороны, простота добавления модификации. Поэтому полезным является представление формул (1.6) в матричной форме, что позволяет использовать эффективные программные пакеты для матричных вычислений.

Через s_{dw} обозначается выражение $\sum_t \varphi_{wt}\theta_{td}$, фактически, это предсказание для вероятности $p(w|d)$. Тогда

$$p_{tdw} = \frac{\varphi_{wt}\theta_{td}}{\sum_t \varphi_{wt}\theta_{td}} = \frac{\varphi_{wt}\theta_{td}}{s_{dw}}.$$

Подставляя это выражение в n_{wt} получается

$$n_{wt} = \sum_d n_{dw} \frac{\varphi_{wt}\theta_{td}}{s_{dw}} = \varphi_{wt} \sum_d \theta_{td} \frac{n_{dw}}{s_{dw}}.$$

Аналогично,

$$n_{td} = \theta_{td} \sum_w \varphi_{wt} \frac{n_{dw}}{s_{dw}}.$$

Таким образом, необходимо построить матрицу $\frac{n_{dw}}{s_{dw}}$. Поскольку она является разреженной, то s_{dw} нужно нужно вычислять только для тех d, w, где $n_{dw} > 0$. Эта матрица обозначается за A, она эффективно вычисляется по величинам $n_{dw}, \varphi_{wt}, \theta_{td}$. В этих обозначениях выполнено

$$n_{wt} = \varphi_{wt}(\Theta A)_{tw}, \text{ и } n_{td} = \theta_{td}(A\Phi^T)_{dt}.$$

Перемножение разреженной матрицы A на плотную матрицу Φ^T или Θ выполняется за время $O(|N||T|)$, где $|N|$ — количество ненулевых значений матрицы A, а $|T|$ — вторая размерность соответствующей матрицы.

В случае обобщения оптимизационной задачи (1.7), формулы остаются такими же, только корректируется определение матрицы A:

$$A_{dw} = n_{dw}\ell'(s_{dw}).$$

1.5 Заключение главы

В этой главе были поставлены оптимизационные задачи вероятностного тематического моделирования (1.1) и аддитивной регуляризации тематических моделей (1.4). Для их решения используется EM-алгоритм, состоящий из чередования E-шага (1.2) и M-шага (1.3) и (1.6). Известно, что на практике итерационный алгоритм ARTM всегда сходится. Однако, вопрос теоретического обоснования этого факта не был изучен. Также открытым является вопрос влияния свойств регуляризаторов на сходимость итерационного алгоритма. В следующей главе будут представлены результаты, отвечающие на данные вопросы.

Глава 2. Сходимость алгоритма аддитивной регуляризация тематических моделей

В этой главе будут сформулированы и доказаны основные теоремы о сходимости алгоритма аддитивной регуляризации тематических моделей. Основная идея доказательств будет заключаться в интерпретации алгоритма ARTM как GEM-алгоритма и переиспользовании известных результатов о сходимости GEM-алгоритмов. Также в этой главе будут предложены модификации алгоритма ARTM, улучшающие сходимость и приведены результаты экспериментов на реальных текстовых коллекциях, подтверждающие предложенные улучшения.

2.1 Общие сведения по GEM-алгоритмам

2.1.1 Вероятностные EM- и GEM- алгоритмы

Решается задача максимизации неполного правдоподобия для некой вероятностной модели, в которой есть наблюдаемые переменные X, скрытые переменные Z и параметры Ω:

$$\log p(X \mid \Omega) \to \max_{\Omega}. \tag{2.1}$$

Пусть $q(Z)$ — произвольное распределение на скрытых переменных, тогда:

$$\log p(X \mid \Omega) = \int q(Z) \log p(X \mid \Omega) dZ = \int q(Z) \frac{\log p(X, Z \mid \Omega)}{\log p(Z \mid X,\Omega)} dZ =$$
$$= \int q(Z) \frac{\log p(X, Z \mid \Omega)}{q(Z)} \frac{q(Z)}{\log p(Z \mid X,\Omega)} dZ =$$
$$\underbrace{\int q(Z) \log p(X, Z \mid \Omega) dZ - \int q(Z) \log q(Z) dZ}_{F(q,\Omega)} + \underbrace{\int q(Z) \frac{q(Z)}{p(Z \mid X,\Omega)} dZ}_{KL(q(Z) \,||\, p(Z|X,\Omega))} \tag{2.2}$$

Дивергенция Кульбака-Лейблера $KL(q(Z) \,||\, p(Z \mid X,\Omega))$ оценивает расстояние между двумя распределениями. Основные её свойства:

1. неотрицательность;
2. равна нулю тогда и только тогда, когда распределения совпадают;
3. несимметричность.

В силу неотрицательности KL слагаемое $F(q, \Omega)$ является нижней оценкой на величину $\log p(X \mid \Omega)$. От максимизации $\log p(X \mid \Omega)$ по Ω предлагается перейти к максимизации нижней границы $F(q, \Omega)$ по q и Ω.

EM-алгоритм состоит в итеративном повторении двух шагов:

1. $F(q, \Omega) \to \max\limits_{q}$
2. $F(q, \Omega) \to \max\limits_{\Omega}$

На первом шаге максимизируется выражение

$$F(q, \Omega) \to \max_{q}.$$

Подставляя вместо F его выражение, получается эквивалентное выражение

$$(\log p(X \mid \Omega) - KL(q(Z) \mid\mid p(Z \mid X,\Omega))) \to \max_{q}.$$

$\log p(X \mid \Omega)$ не зависит от q, поэтому выражение эквивалентно

$$KL(q(Z) \mid\mid p(Z \mid X,\Omega)) \to \min_{q}.$$

Из свойств KL-дивергенции следует, что

$$q(Z) = p(Z \mid X,\Omega)$$

То есть на первом шаге необходимо найти или оценить данное апостериорное распределение.

На втором шаге решается задача

$$\operatorname*{argmax}_{\Omega} \left(\int q(Z) \log p(X, Z \mid \Omega)\ dZ - \int q(Z) \log q(Z)\ dZ \right) =$$

$$= \operatorname*{argmax}_{\Omega} \int q(Z) \log p(X, Z \mid \Omega)\ dZ = \mathbf{E}_{q(Z)} \log p(X, Z \mid \Omega).$$

Таким образом, EM-алгоритм заключается в чередовании двух шагов. E (Expectation) соответствует подготовке к вычислению математического ожидания; M (Maximization) — максимизация математического ожидания логарифма правдоподобия по параметрам.

$$\textbf{E-шаг: } \operatorname*{argmin}_{q(Z)} KL(q(Z) \,||\, p(Z \mid X,\Omega)) = p(Z \mid X,\Omega). \tag{2.3}$$

$$\textbf{M-шаг: } \mathbf{E}_{q(Z)} \log p(X, Z \mid \Omega) \to \max_{\Omega}. \tag{2.4}$$

На каждом из этих шагов возникают определённые трудности. Может оказаться, что апостериорное распределение на скрытых переменных невозможно точно найти, поэтому используют приближённые методы (сэмплирование Гиббса) или ищут наиболее подходящее распределение в некотором классе (Variational Bayes). На втором шаге может оказаться, что нельзя найти точную точку максимума функций, поэтому ставится задача не максимизировать, но увеличить значение функционала по сравнению с Ω на предыдущей итерации. Такой подход называют Generalized Expectation Maximization (GEM) алгоритмом.

Пусть теперь стоит задача максимизации не апостериорной вероятности $p(X \mid \Omega)$, а максимизация полной вероятности $p(X, \Omega)$, учитывая некоторую априорную информацию о модели $p(\Omega)$. По формуле условной вероятности $p(X, \Omega) = p(X \mid \Omega)\ p(\Omega)$, повторяя старую декомпозицию(2.2), получаем оптимизационную задачу:

$$\begin{gathered} \log p(X \mid \Omega) \to \max_{\Omega} \\ F(q, \Omega) + KL(q(Z) \,||\, p(Z \mid X,\Omega)) + \log p(\Omega) \to \max_{\Omega} \end{gathered} \tag{2.5}$$

При максимизации $F(q, \Omega) + \log p(\Omega)$ по q и Ω:

E-шаг остаётся без изменений, так как новое слагаемое не зависит от q.

M-шаг меняется соответствующе: $\mathbf{E}_{q(Z)} \log p(X, Z \mid \Omega) + \log p(\Omega) \to \max_{\Omega}$ Алгоритм ARTM имеет в точности такой вид, если интерпретировать R как $\log p(\Omega)$, хотя формально для вывода не нужна вероятностная природа для $p(\Omega)$, поскольку он участвует только в оптимизационной задаче для M-шага.

2.1.2 Известные результаты о сходимости

В работе [36] представлены общие результаты о сходимости EM- и GEM-алгоритмов, предложенных в работе [35]. Базовой теоремой, с помощью которой

доказываются сходимости является Global Convergence Theorem, предложенная в работе [46].

Определение 1. *Отображение $A\colon X \to 2^X$ будем называть point-to-set отображением на множестве X. Такое отображение называется замкнутым, если из $x_k \to x$, $x \in X$, $y_k \to y$ и $y_k \in A(x_k)$ следует, что $y \in A(x)$.*

Теорема 3 (Global Convergence Theorem)**.** *Пусть $\{x_k\}_{k=0}^{\infty}$, $x_k \in X$ — последовательность, порождённая правилом $x_{k+1} \in M(x_k)$, где M — point-to-set отображение на множестве X. Пусть дано некоторое множество решений $\Gamma \subset X$, а также*

1. *Все x_k принадлежат некоторому компакту $S \subset X$.*
2. *M – замкнуто на $X \setminus \Gamma$.*
3. *Существует непрерывная функция α такая, что, во-первых, при $x \notin \Gamma$ выполнено $\alpha(y) > \alpha(x)$ для любого $y \in M(x)$, и, во-вторых, при $x \in \Gamma$ выполнено $\alpha(y) \geqslant \alpha(x)$ для любого $y \in M(x)$.*

Тогда все предельные точки $\{x_k\}_{k=0}^{\infty}$ находятся в множестве Γ и $\alpha(x_k)$ монотонно сходится к $\alpha(x^*)$ для некоторого $x^* \in \Gamma$.

Чтобы сформулировать результаты для GEM-алгоритмов в терминах данной теоремы, требуется ввести новые обозначения (как в работе [36]).

Множество Ω — множество параметров модели, в котором ведётся оптимизация. Совпадает с множеством X из теоремы. Элемент множества Ω обозначается φ. Максимизируемая функция логарифма правдоподобия (2.1) при некотором значении параметров модели φ обозначается $L(\varphi)$. Соответственно оптимизационная задача записывается как $\max\limits_{\varphi \in \Omega} L(\varphi)$

На E-шаге (2.3) определяется некоторое распределение при определённом наборе параметров $\varphi \in \Omega$. По этому распределению считается математическое ожидание $\mathbf{E}_{p(Z|X,\varphi)} \log p(X, Z \mid \psi)$ при некотором наборе параметров $\psi \in \Omega$, это выражение обозначается за $Q(\psi, \varphi)$. Таким образом на M-шаге (2.4) происходит максимизация (в случае EM-алгоритма) или увеличение (в случае GEM-алгоритма) функции $Q(\psi, \varphi)$ по $\psi \in \Omega$, а φ берётся с предыдущей итерации. Преобразование φ в точку максимизации (или увеличения) Q задаётся point-to-set отображение M, где значением является множество из одной точки.

В качестве множества решений Γ берётся множество

$$\mathcal{L} = \text{множество стационарных точек } L,$$

либо множество

$$\mathcal{M} = \text{множество локальных максимумов } L.$$

Применение Теоремы 3 к GEM-алгоритму позволяет получить следующее утверждение

Теорема 4. *Пусть* $\{\varphi_k\}_{k=0}^{\infty}$, $\varphi_k \in \Omega$ — *последовательность, порождённая правилом* $\varphi_{k+1} \in M(\varphi_k)$. *Пусть*

1. M – *замкнуто на* $\Omega \setminus \mathcal{L}$*(соотв.* $\mathcal{M}$*).*
2. $L(\varphi_{k+1}) > L(\varphi_k)$ *для всех* $\varphi_k \notin \mathcal{L}$*(соотв.* $\mathcal{M}$*).*

Тогда все предельные точки $\{\varphi_k\}_{k=0}^{\infty}$ *находятся в множестве* $\mathcal{L}$*(соотв.* $\mathcal{M}$*) и* $L(\varphi_k)$ *монотонно сходится к* $L(\varphi^*)$ *для некоторого* $\varphi^* \in \mathcal{L}$*(соотв.* $\mathcal{M}$*).*

Если функционал $Q(\psi, \varphi)$ непрерывен по ψ и θ, то этого достаточно для замкнутости M. Второе же условие нельзя гарантировать в общем виде для GEM-алгоритмов и оно является предметом доказательства.

Теорема 4 не гарантирует сходимости параметров φ_k, она гарантирует слабую сходимость в смысле функционала L. Для многих GEM-алгоритмов может быть доказано, что

$$||\varphi_{k+1} - \varphi_k|| \to 0 \text{ при } k \to \infty.$$

В этих ограничениях теорему о сходимости GEM-алгоритма можно усилить:

Определение 2. *Пусть* L — *некоторая функция,* S — *подмножество области определения* L, *а* x — *некоторое значение из* $\mathcal{R}$. *За* $\eta_f L(S, x)$ *обозначим*

$$\{\varphi\colon \varphi \in A \text{ и } fL(\varphi) = a\}.$$

Теорема 5. *Пусть* $\{\varphi_k\}_{k=0}^{\infty}$, $\varphi_k \in \Omega$ — *GEM-последовательность, удовлетворяющая условиям Теоремы 4. Пусть также выполняется*

$$||\varphi_{k+1} - \varphi_k|| \to 0 \text{ при } k \to \infty.$$

Тогда все предельные точки $\{\varphi_k\}_{k=0}^{\infty}$ *находятся в связном и компактном подмножестве* $\eta_L(\mathcal{L}, L^*)$*(соотв.* $\eta_L(\mathcal{M}, L^*)$*), где* L^* *это предел* $L(\varphi_k)$. *Если же множество* $\eta_L(\mathcal{L}, L^*)$*(соотв.* $\eta_L(\mathcal{M}, L^*)$*) дискретно, то есть все его связанные подмножества являются одноэлементными, то* φ_k *сходится к некоторому* $\varphi^* \in \eta_L(\mathcal{L}, L^*)$*(соотв.* $\eta_L(\mathcal{M}, L^*)$*).*

Таким образом, для доказательства сходимости GEM-алгоритма требуется показать увеличение L на каждой итерации, а также стремление разности параметров на соседних итерациях к нулю.

2.1.3 EM-алгоритм максимизации неполного правдоподобия в модели PLSA

Модель PLSA уже была введена в главе 1, однако, рассматривалось дискретное вероятностное пространство на $D \times W \times T$. Модель возможно задать иным способом, чтобы проделать вероятностный вывод EM алгоритма. Для этого расширяется вероятностное пространство. С каждой словопозицией слова в документе d связывается одна определенная тема t. За Z обозначаются темы всех словопозиций (d, i) в коллекции. За w_{di} обозначим i-ое слово в документе d, а за z_{di} его тему. Тогда расширенную вероятность можно записать следующим образом:

$$p(D, Z \mid \Phi, \Theta) = \prod_{d \in D} \prod_{i=1}^{N_d} p(w_{di}, z_{di} \mid \Phi, \Theta) = \prod_{d \in D} \prod_{i=1}^{N_d} \varphi_{w_{di} z_{di}} \theta_{z_{di} d}.$$

Поскольку темы — это ненаблюдаемые величины, то данную модель факторизуют по Z, получая

$$P(D \mid \Phi, \Theta) = \sum_{Z} p(D, Z \mid \Phi, \Theta),$$

и максимизируют данное выражение по Φ и Θ.

На E-шаге необходимо оценить распределение на скрытых переменных при условии параметров и наблюдаемых величин: $p(Z \mid X,\Phi,\Theta)$. Так как словопозиции независимы, то сразу можно перейти к отдельным вероятностям:

$$p(Z \mid D,\Phi,\Theta) = \prod_{d \in D} \prod_{i=1}^{N_d} p(z_{di} \mid w_{di}, \Phi, \Theta)$$

Эти вероятности находятся по формуле Байеса:

$$p(z_{di} \mid w_{di}, \Phi, \Theta) = \frac{p(w_{di} \mid z_{di}, \Phi, \Theta) p(z_{di} \mid \Phi,\Theta)}{\sum_{t=1}^{T} p(w_{di} \mid t, \Phi, \Theta) p(t \mid \Phi,\Theta)} = \frac{\varphi_{w_{di} z_{di}} \theta_{z_{di} d}}{\sum_{t=1}^{T} \varphi_{w_{di} t} \theta_{td}}$$

Фактически это формула для p_{tdw} из уравнения (1.2).

На M-шаге максимизируется выражение:

$$\mathbf{E}_{p(Z|X,\Phi,\Theta)} \log p(X,Z \mid \Phi,\Theta) = \sum_{d\in D} \sum_{i=1}^{N_d} \mathbf{E}_{p(z_{di}|w_{di},\Phi,\Theta)} (\log \varphi_{w_{di}z_{di}} + \log \theta_{z_{di}d}) \to \max_{\Phi,\Theta}$$

Математическое ожидание проносится внутрь суммы в силу независимости словопозиций, после чего по определению математического ожидания получается:

$$\sum_{d\in D} \sum_{i=1}^{N_d} \sum_{t\in T} p(z_{di} = t \mid w_{di},\Phi,\Theta)(\log \varphi_{w_{di}t} + \log \theta_{td}) \to \max_{\Phi,\Theta}.$$

Причём

$$\sum_{d\in D} \sum_{i=1}^{N_d} \sum_{t\in T} p(z_{di} = t \mid w_{di},\Phi,\Theta)(\log \varphi_{w_{di}t} + \log \theta_{td}) =$$

$$\sum_{d\in D} \sum_{i=1}^{N_d} \sum_{t\in T} p_{tdw_{di}}(\log \varphi_{w_{di}t} + \log \theta_{td}) =$$

$$= \sum_{d\in D} \sum_{w\in W} \sum_{t\in T} n_{dw} p_{tdw} (\log \varphi_{wt} + \log \theta_{td}).$$

При добавлении априорного распределения (2.5) E-шаг не меняются, а M-шаг имеет вид

$$\sum_{d\in D} \sum_{w\in W} \sum_{t\in T} n_{dw} p_{tdw} (\log \varphi_{wt} + \log \theta_{td}) + \log p(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

Полагая $R(\Phi,\Theta) = \log p(\Phi,\Theta)$ и решая данную оптимизационную задачу, получается тот же M-шаг, что и в Теореме 1. E-шаг, как отмечалось ранее, тоже совпадает. Совпадение формул E-шага и M-шага позволяет интерпретировать алгоритм ARTM как GEM-алгоритм и переиспользовать текущие известные результаты о сходимости.

2.2 Теоремы о сходимости алгоритма аддитивной регуляризации тематических моделей

2.2.1 Основная теорема о сходимости

Объединяя Теоремы 4 и 5, и адаптируя обозначения под ARTM, нетрудно получить теорему, с помощью которой удобно доказывать сходимость EM-алгоритма в ARTM.

Теорема 6. *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) *в не стационарных точках* F*. Тогда все предельные точки траектории* (Φ^k, Θ^k) *являются стационарными точками* F*. Если также* $\|\varphi_{wt}^k - \varphi_{wt}^{k+1}\| \to 0$ *и* $\|\theta_{td}^k - \theta_{td}^{k+1}\| \to 0$*, а множество стационарных точек* F *дискретно для каждого уровня значений* F*, то* (Φ^k, Θ^k) *сходится к некоторой стационарной точке* F*.*

Определение 3. *Регуляризатор* R *является* δ-регулярным*, если на итерациях EM-алгоритма* $\forall t\ \exists w\colon n_{wt} + r_{wt} > \delta$ *и* $\forall d\ \exists t\colon n_{td} + r_{td} > \delta$*. 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.

Определение 4. *Регуляризатор* R сохраняет нуль*, если на итерациях алгоритма из* $n_{wt} = 0$ *следует* $\varphi_{wt} = 0$ *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 $\varphi_{wt} > 0$. Однако при использовании ненулевой инициализации φ_{wt} значение n_{wt} не может обратиться в нуль. Поэтому и для такого регуляризатора условие сохранения нуля выполняется.

Определение 5. *Регуляризатор R называется* ε-разреживающим, *если на итерациях EM-алгоритма $\varphi_{wt}, \theta_{td} \notin (0, \varepsilon)$.*

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.

Определение 6. *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 [45]. Он гарантирует, что для любого терма в любом документе найдётся хотя бы одна тема с ненулевой вероятностью.

Введём вспомогательный функционал

$$Q(\Phi, \Theta, \Phi', \Theta') = \sum_{d,w,t} n_{dw} p'_{tdw} \ln(\varphi_{wt}\theta_{td}) + R(\Phi, \Theta), \qquad p'_{tdw} = \frac{\varphi'_{wt}\theta'_{td}}{\sum\limits_t \varphi'_{wt}\theta'_{td}}. \quad (2.6)$$

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 [35] и [36].

Теорема 7. *Пусть регуляризатор R является дифференцируемой функцией при $\varphi_{wt}, \theta_{td} \in (0, 1]$, сохраняющей нуль, корректной, ε-разреживающей и δ-*

регулярной. Также допустим, что $Q(\Phi^{k+1}, \Theta^{k+1}, \Phi^k, \Theta^k) \geqslant Q(\Phi^k, \Theta^k, \Phi^k, \Theta^k)$ *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 $\varphi_{wt}, \theta_{td} \in [\varepsilon, 1]$, 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 М-шаге k-ой итерации были получены матрицы $(\Phi^{k+1}, \Theta^{k+1})$.

По условию теоремы, начиная с некоторой итерации выполнено

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

Подставим сюда вместо Q его выражение по определению:

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

откуда следует

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

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 $\mathrm{KL}\big(p_{tdw}^k \,\big\|\, p_{tdw}^{k+1}\big) \leqslant \Delta(L + R)^k \to 0$ при $n_{dw} > 0$, что завершает доказательство. $\square$

Следствие 1. *Если в дополнение к условиям Теоремы 7 регуляризатор R сильно регулярен, а r_{wt} и r_{td} непрерывны по p_{tdw}, то*

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

Доказательство.
Согласно неравенству Пинскера [47], $\|A - B\|_1 \leqslant 2\sqrt{\mathrm{KL}(A\|B)}$. Поэтому сходимость по KL-дивергенции влечёт за собой сходимость по l_1 норме. φ_{wt} и θ_{td} являются непрерывными функциями от n_{wt}, n_{td}, r_{wt}, r_{td}, которые в свою очередь непрерывно зависят от p_{tdw}. Следовательно, сходимость p_{tdw} влечёт за собой сходимость φ_{wt} и θ_{td}. □

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 для каждого уровня значений F, 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 6 тоже выполнены, если рассматривать все функции на области определения с ограничением на множество ненулевых позиций Ω. □

Таким образом, итерационный процесс EM-алгоритма в ARTM разбивается (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.

2.2.2 Свойства траектории итерационного процесса ARTM

Важным условием в теоремах сходимости является дискретность множества стационарных точек. В силу неединственности стохастического разложения матрицы это условие может не выполняться. Это подводит к поиску альтернативных достаточных условий сходимости. Сходимость итерационного процесса неразрывно связана со свойствами его траектории. Поэтому можно связать свойства траектории процесса с изменениями $L + R$.

Теорема 8. *Пусть выполнены условия Теоремы 7. Тогда сходимость ряда*

$$\sum_{n=1}^{\infty}(\Delta^k L + \Delta^k R)^{\alpha}$$

влечёт за собой сходимость ряда

$$\sum_{n=1}^{\infty}(\Delta^k p_{tdw})^{2\alpha}.$$

Доказательство.
Было доказано, что $KL(p_{tdw}^k||p_{tdw}^{k+1}) \leqslant \Delta(L+R)^k$. По неравенству Пинскера [47]

$$||p_{dw}^k - p_{dw}^{k+1}||_1 \leqslant C \cdot \sqrt{KL(p_{tdw}^k||p_{tdw}^{k+1})} \leqslant C\sqrt{\Delta^k(L+R)},$$

а, значит, $(\Delta^k p_{tdw})^2 \leqslant C^2\Delta^k(L+R)$, откуда очевидно следует требуемое утверждение. □

Следствие 3. *В условиях теоремы 7 ряд* $\sum\limits_{n=1}^{\infty}(\Delta p_{tdw}^k)^{2\alpha}$ *сходится при* $\alpha \geqslant 1$.

Доказательство.
Монотонность по α свойства сходимости очевидна. При $\alpha = 1$ имеем

$$\sum_{n=1}^{m}(\Delta^k L + \Delta^k R) = (L^{(m)} + R^{(m)}) - (L^{(0)} + R^{(0)})$$

А сходимость данной последовательности уже была доказана. □

Следствие 4. *В условиях теоремы 7 условие дискретности множества стационарных точек можно заменить условием сходимости ряда*

$$\sum_{n=1}^{\infty}\sqrt{\Delta^k L + \Delta^k R}.$$

К сожалению, это неконструктивное условие. Однако, стоит принять во внимание, что при машинных вычислениях, начиная с некоторого момента, изменения функционалов меньше машинной точности, и к этому моменту на практике частичная сумма ряда не уходит в бесконечность. Поэтому на реальных коллекциях этот ряд вычислительно сходится.

2.2.3 Эксперимент по проверке достаточных условий теоремы о сходимости

Два основных условия Теоремы 7, выполнение которых на реальной коллекции нужно проверить — это сохранение нуля и ε-разреживание. Для экспериментальной этих условий мы использовали лемматизированную коллекцию новостных сообщений на английском языке «20 NewsGroups» [48]. Тематическая модель строилась EM-алгоритмом для ARTM, описанным в [45], с использованием регуляризатора декоррелирования [49]:

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

и регуляризаторами разреживания (1.5).

Эти регуляризаторы были выбраны как одни из наиболее часто используемых. Регуляризаторы разреживания зануляют являются неограниченными сверху и именно из-за них возникает необходимость вводить требование ε-разреживания для регуляризаторов. Максимизация регуляризатора декоррелирования 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.

Измерялись две основные метрики: минимальное ненулевое значение матриц Φ и Θ, а также доля нулевых значений в этих же матрицах.

Рисунки 2.1 и 2.2 показывают, что ожидаемо с ростом коэффициента регуляризации увеличивается степень разреженности, но также, что разреженность монотонна на итерациях и постепенно выходит на плато. Это согласуется с интерпретацией итераций Теоремой 7, согласно которой сначала происходит

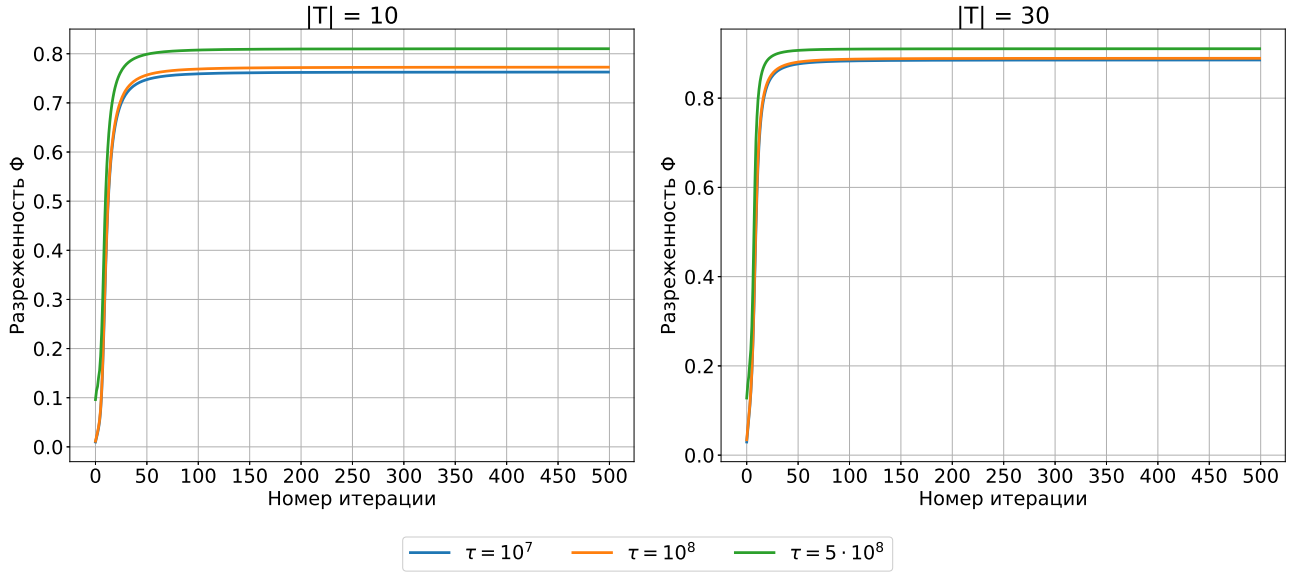

Рисунок 2.1 — Доля нулевых элементов в матрице Φ на итерациях, при различных значениях коэффициента регуляризации τ.

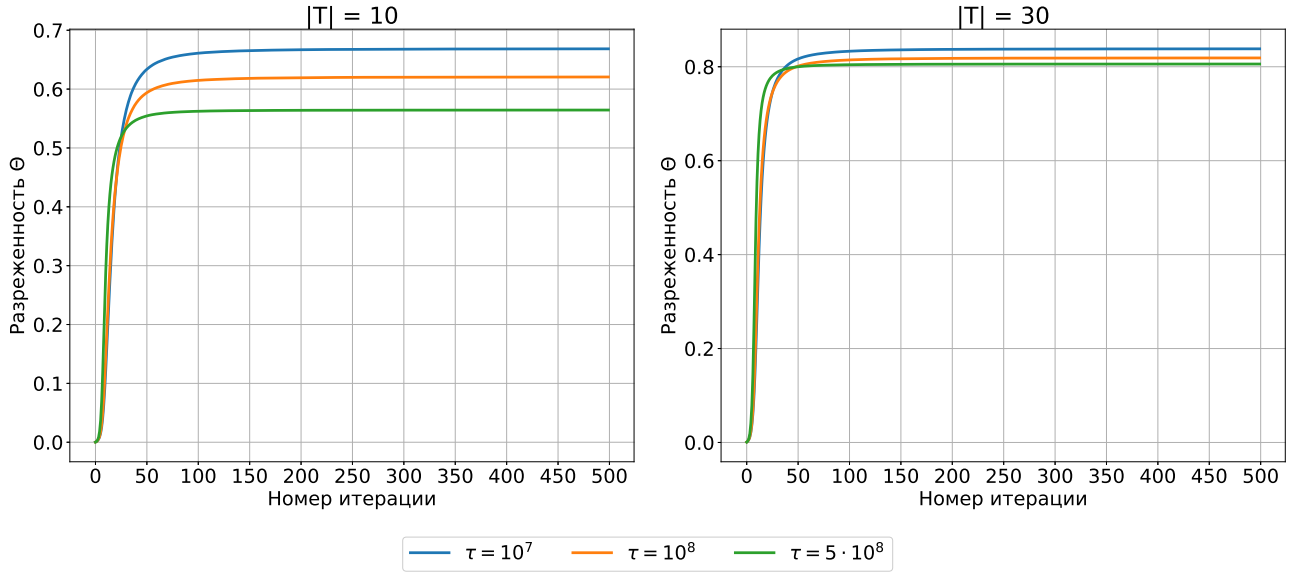

Рисунок 2.2 — Доля нулевых элементов в матрице Θ на итерациях, при различных значениях коэффициента регуляризации τ.

стабилизация множества нулей матриц Φ и Θ, а затем значение функционала дооптимизируется.

Рисунки 2.3 и 2.4 показывают, что минимальное ненулевое значение на первых итерациях уменьшается, затем достигает своего минимума и дальше продолжает расти. Уменьшение на первых итерациях происходит из-за случайной начальной инициализации матриц. Этот результат также подтверждает выполнение условия ε-разреживания для регуляризатора.

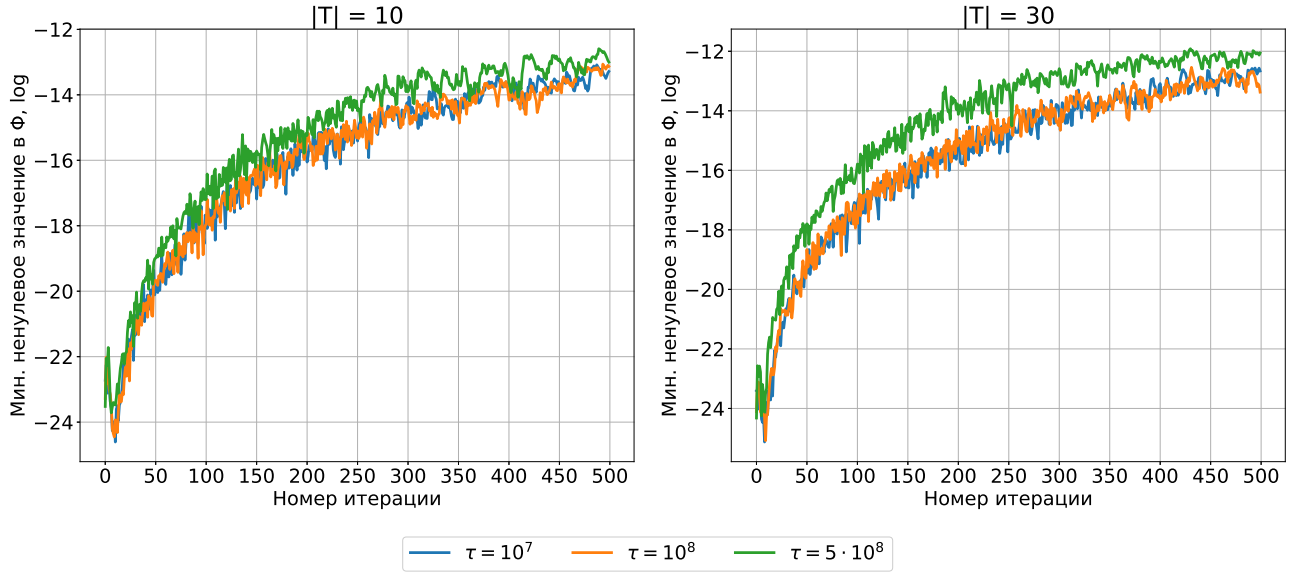

Рисунок 2.3 — Минимальное ненулевое значение в матрице Φ на итерациях, при различных значениях коэффициента регуляризации τ.

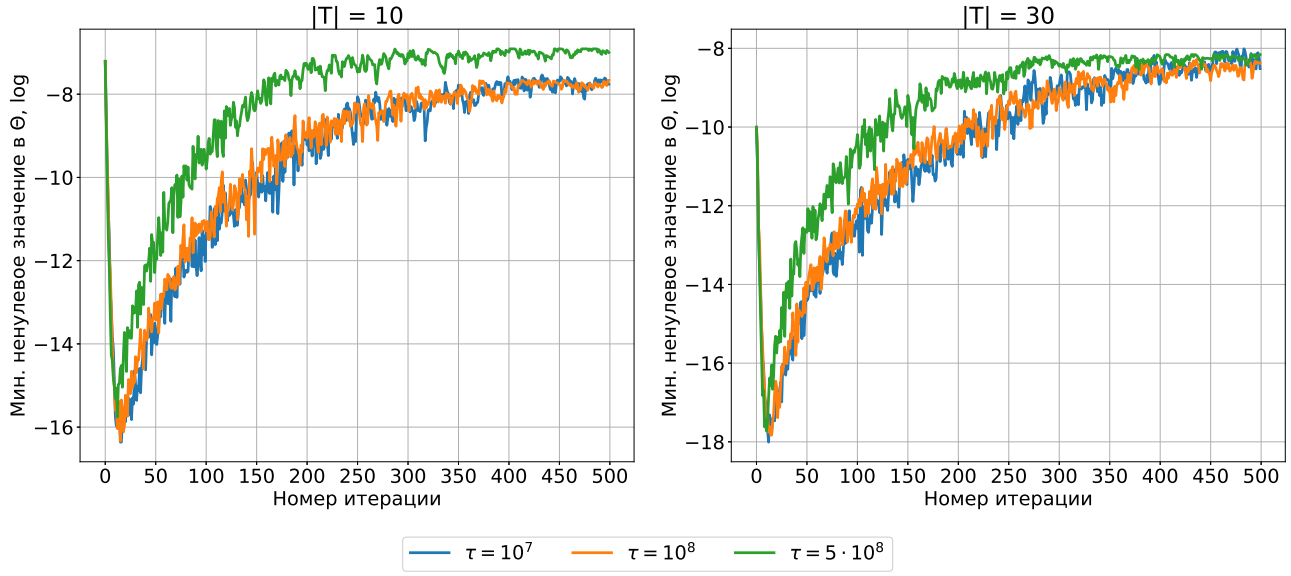

Рисунок 2.4 — Минимальное ненулевое значение в матрице Θ на итерациях, при различных значениях коэффициента регуляризации τ.

2.3 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

$$\bar{Q}(\Phi, \Theta, \Phi', \Theta') = \sum_{d,w,t} n_{dw} p'_{tdw} \ln(\varphi_{wt}\theta_{td}).$$

Тогда $Q = \bar{Q} + R$.

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

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

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

$$\begin{cases} \varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(n_{wt} + r_{wt}\right), \\ \theta_{td} = \underset{t \in T}{\operatorname{norm}}\left(n_{td} + r_{td}\right) \end{cases} \tag{2.7}$$

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:

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

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

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:*

$$g_{wt} = \frac{1}{m_t} \sum_{u \in W} \left(\frac{\partial R}{\partial \varphi_{wt}} - \frac{\partial R}{\partial \varphi_{ut}} \right) \varphi_{ut},$$
$$g_{td} = \frac{1}{m_d} \sum_{s \in T} \left(\frac{\partial R}{\partial \theta_{td}} - \frac{\partial R}{\partial \theta_{sd}} \right) \theta_{sd}.$$

Доказательство.
В силу нормировки

$$\varphi_{wt} = \frac{m_{wt}}{\sum_w m_{wt}}.$$

Отсюда

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

Следовательно,

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

Формула для g_{dt} доказывается аналогично. □

Теперь докажем основную теорему раздела.

Теорема 9. *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 (2.7) без занулений элементов матриц угол между вектором изменений Δn и градиентом $\bar{R}$ острый, если градиент ненулевой.

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

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

В силу симметрии суммы выполнено:

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

Пусть здесь достигается равенство, тогда $\frac{\partial R}{\partial\varphi_{wt}}=\frac{\partial R}{\partial\varphi_{ut}}$ для всех u и w, где $\varphi_{wt}>0$ и $\varphi_{ut}>0$. Тогда

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

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

Ранее было показано (Теорема 7), 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.

Также стоит отметить, что если регуляризационные поправки вычисляются по n_{wt} и n_{td}, то будут непрерывными функциями от p_{tdw}, что важно для сходимости параметров.

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.

2.3.1 Стремление коэффициента регуляризатора к нулю

Один из рычагов управления регуляризацией в алгоритме ARTM — это изменение коэффициента регуляризации. Для удобства будем считать, что регуляризатор на k-ой итерации алгоритма равен $\tau^k R$, а верхний индекс у функционалов означает, что они рассчитываются на соответствующей итерации. При анализе регуляризационного преобразования использовался подход с подсчётом градиента не по φ_{wt}, а по n_{wt}. Используя данный подход, докажем следующее утверждения для коэффициентов регуляризации, стремящихся к нулю.

Утверждение 2. *Существует такая константа γ, что если $\tau^k \leqslant \gamma \Delta \bar{Q}^k$, а также $\dfrac{1}{n_t}\dfrac{\partial \bar{R}}{\partial \varphi_{wt}}(n_{wt}, n_{td})$ — ограниченная функция (константой C), то при*

$$r_{wt} = \tau \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}} \left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d} \right)$$

и

$$r_{td} = \tau \frac{n_{td}}{n_d} \frac{\partial R}{\partial \theta_{td}} \left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d} \right)$$

будет выполнено $\Delta \bar{Q}^k \geqslant 0$.

Доказательство.
Для лаконичности рассмотрим случай $R(\Phi, \Theta) = R(\Phi)$.
При регуляризационном преобразовании выполнено

$$\Delta^k n_{wt} = \left(n_{wt} + \tau^k \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}} \right)_+ - n_{wt}.$$

Поэтому

$$\left(n_{wt} + \tau^k \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right)_+ - n_{wt} \leqslant n_{wt} + \left|\tau^k \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right| - n_{wt} \leqslant \left|\tau^k \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right|,$$

$$\left(n_{wt} + \tau^k \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right)_+ - n_{wt} \geqslant n_{wt} - \left|\tau^k \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right| - n_{wt} \geqslant -\left|\tau^k \frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right|,$$

$$|\Delta^k n_{wt}| \leqslant \tau^k \left|\frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right|.$$

Подставим значение градиента из Утверждения 1:

$$\left|\langle \Delta^k n, \nabla R(n_{wt}, n_{td})\rangle\right| = \left|\sum_{w,t,u} \frac{1}{n_t}\left(\frac{\partial R}{\partial \varphi_{wt}} - \frac{\partial R}{\partial \varphi_{ut}}\right) \Delta^k n_{wt} \frac{n_{ut}}{n_t}\right| \leqslant$$

$$\leqslant \sum_{w,t,u} 2C \left|\Delta^k n_{wt}\right| \left|\frac{n_{ut}}{n_t}\right| \leqslant \sum_{w,t,u} 2C\tau^k \left|\frac{n_{wt}}{n_t} \frac{\partial R}{\partial \varphi_{wt}}\right| \left|\frac{n_{ut}}{n_t}\right| \leqslant$$

$$\leqslant \sum_{w,t,u} 2C^2 \tau^k n_{wt} \left|\frac{n_{ut}}{n_t}\right| \leqslant \sum_{w,t} 2C^2 n_{wt} \leqslant$$

$$\leqslant 2C^2 \sum_w n_w \tau^k \leqslant (2\gamma C^2 \sum_w n_w) \Delta \bar{Q}^k.$$

Если $2\gamma C^2 \sum_w n_w < 1$, то изменение $\bar{Q}^k$ на этапе регуляризационного преобразования меньше чем на этапе максимизации $\bar{Q}^k$, а, значит, суммарный эффект будет положительным. □

2.4 Классификация регуляризаторов

С точки зрения изменения функционала Q стоит выделить несколько типов регуляризаторов.

Аналитические регуляризаторы. В эту группу попадают регуляризаторы, для которых возможно явно найти решение задачи максимизации Q на М-шаге. В этом случае не требуется анализировать углы между градиентами, увеличение функционала получается по построению. Таковыми регуляризаторами являются, например, регуляризаторы сглаживания и разреживания. Аналитические регуляризаторы обладают ещё одним важным свойством: их воздействие можно считать отдельно. Пусть $R = R_1 + R_2$, где R_1 — аналитический

регуляризатор. На M-шаге необходимо построить увеличение функционала $\bar{Q} + R = \bar{Q} + R_1 + R_2$. По формулам M-шага вычисляется (n_{wt}, n_{td}) как точка максимума Q, а затем увеличивается R. Однако, можно определить (n_{wt},n_{td}) как точку максимума $\bar{Q} + R_1$ (это можно сделать в силу аналитичности R_1), а затем производить увеличение R_2. Таким образом, численные методы оптимизации будут использоваться только для той части регуляризатора, где не получается явно найти максимум.

Вогнутые регуляризаторы. Если R вогнутая функция, то $\bar{Q} + R$ тоже вогнутая функция, и, следовательно, имеет единственный максимум. При некоторых дополнительных допущениях будет происходить увеличение $\bar{Q} + R$. Однако, в случае вогнутого регуляризатора можно сказать, что на шаге регуляризационного преобразования происходит приближение к глобальному максимуму, а не просто увеличение значения. Таковыми регуляризаторами являются регуляризаторы когерентности и лапласианы графов связей документов.

Неограниченные регуляризаторы. В случае, если регуляризатор неограничен, задача оптимизации оказывается некорректно поставленной, поскольку максимум оптимизируемой функции равен бесконечности. Однако, за счёт отделения итерационного процесса от нуля, это проблему получается решить (подробнее в Теореме 7).

Произвольные регуляризаторы. Для произвольных регуляризаторов было доказано увеличение R при регуляризационном преобразовании (Теорема 9). При дополнительных условиях оно преобразуется в увеличение $\bar{Q} + R$ на итерациях.

2.5 Модификация M-шага алгоритма ARTM

Обычно в реализациях EM-алгоритма для ARTM [30; 31; 50] 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 9 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 M-шаге k-й итерации, то будут выполнены теоретические гарантии оптимальности. На основе этой Теоремы предлагается модифицировать M-шаг алгоритма.

2.5.1 Описание модификации

Обычные формулы M-шага для регуляризационных поправок

$$r^k_{wt} = \varphi^{k-1}_{wt}\frac{\partial R}{\partial \varphi_{wt}}\big(\Phi^{k-1}_{wt},\Theta^{k-1}_{td}\big); \qquad r^k_{td} = \theta^{k-1}_{td}\frac{\partial R}{\partial \theta_{td}}\big(\Phi^{k-1}_{wt},\Theta^{k-1}_{td}\big); \tag{2.8}$$

заменяются на модифицированные согласно Теореме 9:

$$r^k_{wt} = \frac{n^k_{wt}}{n^k_t}\frac{\partial R}{\partial \varphi_{wt}}\left(\frac{n^k_{wt}}{n^k_t},\frac{n^k_{td}}{n^k_d}\right); \qquad r^k_{td} = \frac{n^k_{td}}{n^k_d}\frac{\partial R}{\partial \theta_{wt}}\left(\frac{n^k_{wt}}{n^k_t},\frac{n^k_{td}}{n^k_d}\right). \tag{2.9}$$

Рассмотрим поведение двух формул M-шага на примере простого регуляризатора. Пусть $R = -\tau\sum\limits_{w,t}\varphi_{wt}$. Формально, он не должен влиять на оптимизацию, поскольку равен константе при ограничениях задачи. Однако, стандартные формулы дадут следующий M-шаг:

$$\begin{cases} \varphi_{wt} = \underset{w}{\operatorname{norm}}\big(n_{wt} - \tau\varphi_{wt}\big), \\ \theta_{td} = \underset{t}{\operatorname{norm}}\big(n_{td} - \tau\theta_{td}\big). \end{cases}$$

Если не будет занулений, то этот процесс сойдётся, скорее всего, к той же точке, что и PLSA, но траектория будет другой. Используя несмещённые оценки, можно получить:

$$\begin{cases} \varphi_{wt} = \underset{w}{\operatorname{norm}}(\underset{\tau}{\operatorname{sparse}}(n_{wt}, n_t)), \\ \theta_{td} = \underset{t}{\operatorname{norm}}(\underset{\tau}{\operatorname{sparse}}(n_{td}, n_d)), \end{cases}$$

где $\underset{\tau}{\operatorname{sparse}}(x,y) = x$, если $\tau < y$ и 0 иначе. Это уже практически PLSA, но с условием, на селекцию тем: тема должна содержать некоторое минимальное количество слов (параметр n_t), иначе будет удалена.

2.5.2 Эксперимент по оценке эффекта от модификации

Согласно Теореме 9, если рассчитывать регуляризационные поправки по формулам (2.9), 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 функционала.

Для экспериментальной проверки этого утверждение мы как и в разделе 2.2.3 использовали коллекцию «20 NewsGroups» и регуляризатор декоррелирования:

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

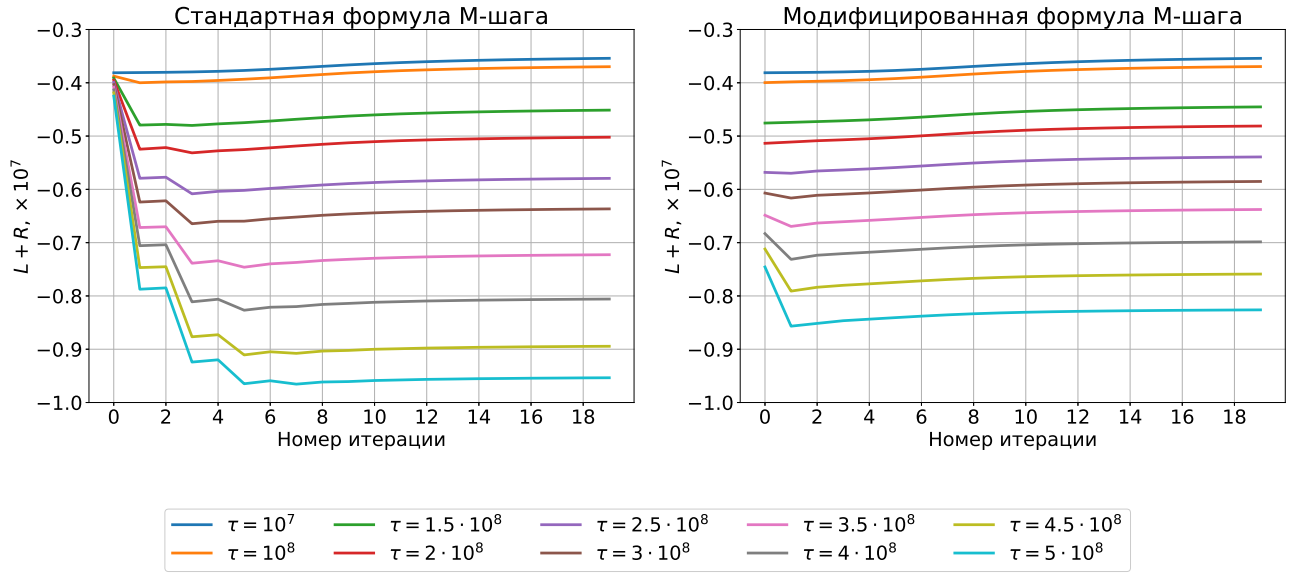

Рисунок 2.5 — Изменение функционала $L + R$ на итерациях, $|T| = 30$, при различных значениях коэффициента регуляризации τ.

τ	$L+R$ стандарт	$L+R$ модификация	Улучшение $L+R$, %
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

Таблица 1 — Итоговые значения $L+R$ по окончании итераций.

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 М-шага: стандартная (2.8) и модифицированная (2.9).

На Рисунке 2.5 видно, что при стандартных формулах М-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 1. Также заметим, что чем больше τ, то есть чем сильнее воздействие регуляризатора на модель, тем существеннее предложенная модификация улучшает полученное решение.

2.6 Обобщение теорем о сходимости на случай общей функции потерь

Результаты Теоремы 7 могут быть перенесены на случай обобщённой оптимизационной задачи (1.7). Это делается в несколько этапов.

2.6.1 Обобщение интерпретации как GEM-алгоритма

Для введения нижних оценок потребуется доказать несколько вспомогательных лемм.

Утверждение 3. *Пусть* $x_t \geqslant 0$ *и* $y_t > 0$, *тогда* $\sum_t x_t - \sum_t y_t \geqslant \sum_t y_t \log \frac{x_t}{y_t}$.

Доказательство.
Пусть $x_t = y_t + \varepsilon_t$, тогда

$$\sum_t y_t \log \frac{x_t}{y_t} = \sum_t y_t \log\left(1 + \frac{\varepsilon_t}{y_t}\right) \leqslant \sum_t y_t \frac{\varepsilon_t}{y_t} = \sum_t \varepsilon_t = \sum_t x_t - \sum_t y_t.$$

□

Лемма 1. *Пусть* $\ell' \geqslant 0$ *и* $\ell'' \geqslant 0$, *тогда*

$$L(\Phi, \Theta) - L(\Phi^0, \Theta^0) \geqslant \sum_{w,d} n_{dw} \ell' \left(\sum s^0_{tdw}\right) \left(\sum_t s^0_{tdw} \log \frac{s_{tdw}}{s^0_{tdw}}\right),$$

где $s_{tdw} \equiv \varphi_{wt}\theta_{td}$.

Доказательство.

$$L(\Phi, \Theta) - L(\Phi^0, \Theta^0) = \sum_{w,d} n_{dw} \left(\ell\left(\sum_t \varphi_{wt}\theta_{td}\right) - \ell\left(\sum_t \varphi^0_{wt}\theta^0_{td}\right)\right) \geqslant$$

$$\geqslant |\text{в силу } \ell'' \geqslant 0| \geqslant \sum_{w,d} n_{dw} \ell' \left(\sum_t \varphi^0_{wt}\theta^0_{td}\right) \left(\sum_t \varphi_{wt}\theta_{td} - \sum_t \varphi^0_{wt}\theta^0_{td}\right) \equiv$$

$$\equiv \sum_{w,d} n_{dw} \ell' \left(\sum_t s^0_{tdw}\right) \left(\sum_t s_{tdw} - \sum_t s^0_{tdw}\right) \geqslant$$

$$\geqslant |\text{по Утверждению 3 и } \ell' \geqslant 0| \geqslant \sum_{w,d} n_{dw}\ell'\left(\sum_t s^0_{tdw}\right)\left(\sum_t s^0_{tdw}\log\frac{s_{tdw}}{s^0_{tdw}}\right).$$

□

По аналогии с (2.6) введём функционал

$$Q(\Phi,\Theta,\overline{\Phi},\overline{\Theta}) = \sum_{w,d} n_{dw}\ell'\left(\sum_t \overline{s}_{tdw}\right)\sum_t \overline{s}_{tdw}\log s_{tdw} + R(\Phi,\Theta) \tag{2.10}$$

Из Леммы 1 следует, что

$$L(\Phi,\Theta)+R(\Phi,\Theta)-L(\Phi^0,\Theta^0)-R(\Phi^0,\Theta^0) \geqslant Q(\Phi,\Theta,\Phi^0,\Theta^0)-Q(\Phi^0,\Theta^0,\Phi^0,\Theta^0).$$

Таким образом, по аналогии с GEM-алгоритмом для роста правдоподобия на каждой итерации нужно обеспечить увеличение $\mathrm{Q}(\Phi,\Theta,\Phi^0,\Theta^0)$ по сравнению с $Q(\Phi^0,\Theta^0,\Phi^0,\Theta^0)$:

$$\sum_{w,d} n_{dw}\ell'\left(\sum_t s^0_{tdw}\right)\left(\sum_t s^0_{tdw}\log\frac{s_{tdw}}{s^0_{tdw}}\right) + R(\Phi,\Theta) - R(\Phi^0,\Theta^0) \geqslant 0$$

для этого можно максимизировать выражение

$$\sum_{w,d} n_{dw}\ell'\left(\sum_r s^0_{rdw}\right)\left(\sum_t s^0_{tdw}\log s_{tdw}\right) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

Если подставить вместо s_{tdw} его выражение и упростить, то получится

$$Q(\Phi,\Theta,\overline{\Phi},\overline{\Theta}) = \sum_{w,d} n_{dw}\ell'\left(\sum_r \overline{s}_{rdw}\right)\left(\sum_t \overline{s}_{tdw}\log s_{tdw}\right) + R(\Phi,\Theta) =$$

$$= \sum_{w,d} n_{dw}\sum_t\left(\overline{\varphi}_{wt}\overline{\theta}_{td}\ \ell'\left(\sum_r \overline{\varphi}_{wr}\overline{\theta}_{rd}\right)\right)(\log\varphi_{wt}+\log\theta_{td}) + R(\Phi,\Theta) =$$

$$= \sum_{w,d,t} n_{dw}\overline{p}_{tdw}\,(\log\varphi_{wt}+\log\theta_{td}) + R(\Phi,\Theta).$$

Поскольку в обозначениях Теоремы 2 выполнено $p^0_{tdw} = \varphi^0_{wt}\theta^0_{td}\ell'\left(p(w|d)\right)$. Таким образом, валидно интерпретация решений систем уравнений Теоремы 2 методом простых итераций как GEM-алгоритма.

2.6.2 Сходимость параметров алгоритма

Если $L+R$ — ограниченная сверху функция, то последовательность $L_n + R_n$ сходится, так как она ограничена сверху и монотонно возрастает. Значит, $\Delta(L+R) \to 0$, следовательно

$$\sum_{w,d} n_{dw} \ell' \left(\sum s^0_{tdw} \right) \left(\sum_t s_{tdw} - \sum_t s^0_{tdw} \right) + \Delta R \to 0 \tag{2.11}$$

и

$$\sum_{w,d} n_{dw} \ell' \left(\sum s^0_{tdw} \right) \left(\sum_t s^0_{tdw} \log \frac{s_{tdw}}{s^0_{tdw}} \right) + \Delta R \to 0. \tag{2.12}$$

То есть имеет место быть слабая сходимость параметров s_{tdw}. Но обычная сходимость s_{tdw} отсюда не следует. Для решения этой проблемы, нужно немного усложнить анализ Леммы 1.

Утверждение 4. *Пусть* $\ell(x) = h(\log(x))$, $\ell'(x) \geqslant 0$ *и* $\ell''(x) \geqslant 0$ *при* $x > 0$, *тогда* $h'(\log x) \geqslant 0$ *и* $h''(\log x) \geqslant 0$ *при* $x > 0$.

Доказательство.

Пусть $\ell(x) = h(\log x)$, тогда $\ell'(x) = \frac{h'(\log x)}{x}$. А вторая производная $\ell''(x) = \frac{h''(\log x) - h'(\log x)}{x^2}$. Отсюда $h''(\log x) = x^2 \ell''(x) + h'(\log x)$.

Так как $x > 0$, то $(\ell''(x) \geqslant 0 \ \wedge \ \ell'(x) \geqslant 0) \to (h'(\log x) \geqslant 0 \ \wedge \ h''(\log x) \geqslant 0)$. То есть, $h'(\log x) \geqslant 0 \ \wedge \ h''(\log x) \geqslant 0$. □

Лемма 2. *Пусть* $\ell(x) = h(\log x)$, $h'(\log x) \geqslant 0$ *и* $h''(\log x) \geqslant 0$ *при* $x > 0$, *тогда*

$$L(\Phi, \Theta) - L(\Phi^0, \Theta^0) \geqslant \sum_{w,d} n_{dw} \ell' \left(\sum_t s^0_{tdw} \right) \left(\sum_t s^0_{tdw} \log \frac{s_{tdw}}{s^0_{tdw}} \right) +$$

$$+ \sum_{w,d} n_{dw} h' \left(\log \sum_t s^0_{tdw} \right) KL \left(\frac{s^0_{tdw}}{\sum_t s^0_{tdw}} \,||\, \frac{s_{tdw}}{\sum_t s_{tdw}} \right).$$

Доказательство.

$$L(\Phi, \Theta) - L(\Phi^0, \Theta^0) = \sum_{w,d} n_{dw} \left(l \left(\sum_t s_{tdw} \right) - l \left(\sum_t s^0_{tdw} \right) \right) =$$

$$= \sum_{w,d} n_{dw} \left(h\left(\log \sum_t s_{tdw} \right) - h\left(\log \sum_t s^0_{tdw} \right) \right) \geqslant |\text{в силу } \ell'' \geqslant 0| \geqslant$$

$$\geqslant \sum_{w,d} n_{dw} h'\left(\log \sum_t s^0_{tdw} \right) \left(\log \sum_t s_{tdw} - \log \sum_t s^0_{tdw} \right) =$$

$$= \sum_{w,d} n_{dw} h'\left(\log \sum_t s^0_{tdw} \right) \log \frac{\sum_t s_{tdw}}{\sum_t s^0_{tdw}} \geqslant$$

$$\geqslant \sum_{w,d} n_{dw} h'\left(\log \sum_t s^0_{tdw} \right) \left(-KL\left(\frac{s^0_{tdw}}{\sum_t s^0_{tdw}} \,||\, \frac{s_{tdw}}{\sum_t s_{tdw}} \right) + \log \frac{\sum_t s_{tdw}}{\sum_t s^0_{tdw}} \right) =$$

$$= \sum_{w,d} n_{dw} h'\left(\log \sum_t s^0_{tdw} \right) \left(\sum_t \frac{s^0_{tdw}}{\sum_t s^0_{tdw}} \left(\log \frac{\frac{s_{tdw}}{\sum_t s_{tdw}}}{\frac{s^0_{tdw}}{\sum_t s^0_{tdw}}} + \log \frac{\sum_t s_{tdw}}{\sum_t s^0_{tdw}} \right) \right) =$$

$$= \sum_{w,d} n_{dw} h'\left(\log \sum_t s^0_{tdw} \right) \left(\sum_t \frac{s^0_{tdw}}{\sum_t s^0_{tdw}} \log \frac{s_{tdw}}{s^0_{tdw}} \right) =$$

$$= \sum_{w,d} n_{dw} \ell'\left(\sum_t s^0_{tdw} \right) \left(\sum_t s^0_{tdw} \log \frac{s_{tdw}}{s^0_{tdw}} \right).$$

Отсюда тривиальным образом получается утверждение Леммы. □

Следствие 5. *В условиях Леммы 2 выполнено*

$$L(\Phi,\Theta)+R(\Phi,\Theta)-L(\Phi^0,\Theta^0)-R(\Phi^0,\Theta^0) \geqslant Q(\Phi,\Theta,\Phi^0,\Theta^0)-Q(\Phi^0,\Theta^0,\Phi^0,\Theta^0)+$$

$$+\sum_{w,d} n_{dw} h'\left(\log \sum_t s^0_{tdw} \right) KL\left(\frac{s^0_{tdw}}{\sum_t s^0_{tdw}} \,||\, \frac{s_{tdw}}{\sum_t s_{tdw}} \right).$$

Это означает, что если $\Delta(L+R) \to 0$, то

$$\sum_{w,d} n_{dw} h'\left(\log \sum_t s^0_{tdw} \right) KL\left(\frac{s^0_{tdw}}{\sum_t s^0_{tdw}} \,||\, \frac{s_{tdw}}{\sum_t s_{tdw}} \right) \to 0.$$

Если существует $\gamma > 0$ такое, что на итерациях $h'\left(\log \sum_t s_{tdw}\right) \geqslant \gamma$, то

$$KL\left(\frac{s^0_{tdw}}{\sum_t s^0_{tdw}} \,||\, \frac{s_{tdw}}{\sum_t s_{tdw}} \right) \to 0.$$

2.6.3 Теоремы о сходимости для случая общей функции потерь

Теорема 10. *Пусть регуляризатор R является дифференцируемой функцией при $\varphi_{wt}, \theta_{td} \in (0, 1]$, сохраняющей нуль, корректной, ε-разреживающей и δ-регулярной. Также допустим, что $\ell(x) = h(\log x)$, $h'(\log x) \geqslant 0$ и $h''(\log x) \geqslant 0$ при $x > 0$ и $Q(\Phi^{k+1}, \Theta^{k+1}, \Phi^k, \Theta^k) \geqslant Q(\Phi^k, \Theta^k, \Phi^k, \Theta^k)$ начиная с некоторой итерации k. Тогда последовательность p^k_{tdw} сходится в смысле дивергенции Кульбака–Лейблера для любых d и w таких, что $n_{dw} > 0$:*

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

Доказательство.
Доказательство повторяет доказательство Теоремы 7 с одним дополнением. Как отмечалось в предыдущем разделе, для стремления KL-дивергенции к нулю требуется, чтобы существовало $\gamma > 0$ такое, что на итерациях $h'\left(\log \sum_t \varphi_{wt}\theta_{td}\right) \geqslant \gamma$. Так как регуляризатор ε-разреживающий, а $h''(\log x) \geqslant 0$, то в качестве γ можно взять $h'\left(\log\left(\varepsilon^2 T\right)\right)$. □

Так как доказательства почти совпадают, то и все следствия Теоремы 7 будут верны и для Теоремы 10.

2.7 Заключение главы

Эта глава решает проблему обоснования сходимости алгоритма ARTM при произвольном гладком критерии регуляризации. Полученные ограничения на регуляризатор в Теореме 7 не являются обременительными, легко проверяются и легко обеспечиваются программной реализацией. Также, как показывает Теорема 10, эти результаты могут быть обобщены на случай произвольной функции потерь в оптимизационной задаче (1.7).

Выполнение этих достаточных условий было подтверждено в экспериментах на реальной текстовой коллекции (раздел 2.2.3). Также была предложена модификация M-мага (2.9), которая за счёт теоретических гарантий ускоряет оптимизацию, позволяя за то же число итераций получать лучшие значения

максимизируемого функционала. Улучшение от предложенной модификаций также было подтверждено в эксперименте (раздел 2.5.2).

Таким образом, было показано, что итерационный процесс ARTM сходится, разумным следующим шагом является анализ свойств точки, к которой сошёлся алгоритм. В частности, вопрос о единственности полученного решения. Изучению данной темы будет посвящена следующая глава.

Глава 3. Единственность стохастического матричного разложения

Как было показано в главе 2, итерационный процесс ARTM сходится, и ставится вопрос анализа свойств точки, к которой сошёлся алгоритм. В частности, вопрос о единственности полученного решения. Известно, что оптимизационная задача ARTM имеет неединственное решение, однако, причины этой неединственности подробно не изучались. Эффект неединственности потенциально состоит из двух частей: неединственности точного матричного разложения в точке, в которую сошёлся алгоритм, и мультиэкстремальности оптимизационной задачи. В этой главе определяется степень влияния этих двух факторов. Для этого рассматривается вопрос единственности точного стохастического матричного разложения.

3.1 Общие сведения по стохастическому матричному разложению

В данном разделе будет сформулирована проблема единственности стохастического матричного разложения и приведены результаты основных работ на заданную тему.

3.1.1 Стохастическое матричное разложение

Чтобы определить задачу стохастического матричного разложения, требуется ввести ряд определений.

Определение 7. *Матрица $F \in \mathbb{R}^{n \times m}$ будет называться неотрицательной, если все её элементы неотрицательны.*

Определение 8. *Неотрицательная матрица $F \in \mathbb{R}^{n \times m}$ будет называться стохастической, если* $\forall j \ \sum_i F_{ij} = 1$.

Определение 9. *Пусть дана матрица $F \in \mathbb{R}^{n\times m}$, её неотрицательным (стохастическим) матричным разложением будет называться представление в виде произведения $F = \Phi\Theta$ двух неотрицательных (стохастических) матриц $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$.*

Определение 10. *Пусть дана матрица $F \in \mathbb{R}^{n\times m}$, её матричным разложением полного ранга будет называться представление в виде произведения $F = \Phi\Theta$ двух матриц полного ранга $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$.*

Далее разложением матрицы F называется полноранговое стохастическое матричное разложение, кроме тех случаев, в которых явно сказано, что это стохастическое разложение или неотрицательное разложение. Также предполагается, что в матрице F нет нулевых столбцов.

Если дано разложение $F = \Phi\Theta$ и у матрицы Φ есть хотя бы два различных столбца или у матрицы Θ есть хотя бы две различных строки, то существует другое разложение $F = \Phi S S^{-1}\Theta$, где S — некоторая матрица перестановки. В связи с этим единственность разложения может быть определена с точностью до матрицы перестановки:

Определение 11. *Разложение $F = \Phi\Theta$ будет называться единственным, если для любого другого разложения $F = \Phi'\Theta'$ выполняется $\Phi' = \Phi S, \Theta' = S^{-1}\Theta$, где S — некоторая матрица перестановки.*

Между неотрицательными и стохастическими матричными разложениями, существует взаимосвязь, указанная в работе [37]:

Утверждение 5. *Пусть $F \in \mathbb{R}^{n\times m}$ матрица без нулевых столбцов с неотрицательным разложением $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$. Рассмотрим стохастическую матрицу $\tilde{F}$, $\tilde{F}_{ij} = \frac{F_{ij}}{\sum\limits_i F_{ij}}$. Тогда $\tilde{F} = \tilde{\Phi}\tilde{\Theta}$ является стохастическим разложением матрицы $\tilde{F}$, где*

$$\tilde{\Phi} = \Phi S,$$
$$\tilde{\Theta} = S^{-1}\Theta',$$
$$\Theta' = \frac{\Theta_{ij}}{\sum\limits_i F_{ij}},$$
$$S = \operatorname{diag}((\sum_i \Phi_{i1})^{-1}, \ldots, (\sum_i \Phi_{ik})^{-1}).$$

Таким образом, результаты о единственности неотрицательного матричного разложения могут быть перенесены на единственность стохастического матричного разложения.

3.1.2 Обзор результатов по единственности неотрицательного матричного разложения

Во всех работах, исследующих единственность неотрицательного матричного разложения, формулируются необходимые либо достаточные условия единственности разложения.

В работе [37] вводится геометрическая интерпретация стохастического матричного разложения.

Определение 12. *Стандартным* $(n-1)$*-мерным симплексом называется множество*

$$\Delta_{n-1} = \{x \in \mathbb{R}^n \mid \sum_{i=1}^{n} x^{(i)} = 1, \forall i \; x^{(i)} \geqslant 0\}.$$

Линейной (выпуклой) оболочкой матрицы X называется линейная (выпуклая) оболочка множества её столбцов $\{X_i\}$ и обозначается $\operatorname{span}(X)$ $(\operatorname{conv}(X))$.

Теорема 11. *Пусть дано разложение* $F = \Phi\Theta$, $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$, *тогда*

$$\operatorname{conv}(F) \subset \operatorname{conv}(\Phi) \subset \operatorname{span}(\Phi) \cap \Delta_{n-1}.$$

Разложение $F = \Phi\Theta$ *единственно, тогда и только тогда, когда* $\operatorname{conv}(\Phi)$ *— единственный выпуклый многогранник на* k *вершинах, удовлетворяющий обоим включениям.*

Следствие 6. *Пусть дано разложение* $F = \Phi\Theta$, *пусть также* $\operatorname{conv}(F) = \operatorname{conv}(\Phi)$, *а все вершины* $\operatorname{conv}(\Phi)$ *являются вершинами* $\operatorname{span}(\Phi) \cap \Delta_{n-1}$. *Тогда разложение* $F = \Phi\Theta$ *единственно.*

Данная геометрическая интерпретация в терминах выпуклых многогранников является основой доказательств утверждений о единственности стохастических матричных разложений этой главы. Например, в работе [38] на

основе этой интерпретации также были получены либо необходимые, либо достаточные условия единственности для неотрицательных матричных разложений.

В работе [39] используется геометрическая интерпретация, введённая в работе [37], и доказывается достаточный критерий единственности.

Определение 13. *Паттерном разреженности вектора v называется множество $\{i \mid v_i = 0\}$.*

Например, паттерн разреженности вектора $(4, 0, 0, 2, 0)$ есть $\{1, 2, 4\}$.

Определение 14. *Неотрицательным рангом матрицы F назовем такое минимальное r, что существует неотрицательное разложение матрицы $F = \Phi\Theta$ ранга r.*

Определение 15. *Пусть дан выпуклый многогранник M, а $v \in M$ — точка многогранника. Тогда v будет называться вершиной M, если для любого отрезка $[u_1, u_2]$, содержащего v, не выполняется $[u_1, u_2] \subset M$.*

Теорема 12. *Пусть дана стохастическая матрица F с рангом, совпадающим с неотрицательным рангом и равным k. Если у матрицы F имеется k ненулевых столбцов таких, что в каждом их них есть $k-1$ нулей и соответствующие этим нулям строки имеют различные паттерны разреженности, тогда матрица F имеет единственное разложение.*

Также в работе описывается техника предобработки данных, приводящая к устойчивости и разреженности получаемой тематической модели. Демонстрируется эффективность техники на нескольких наборах изображений.

3.2 Теорема о единственности разложения

В этом разделе будет сформулирована и доказана основная теорема главы о достаточных условиях единственности стохастического матричного разложения. Основой доказательства является лемма условиях, при которых точка является вершиной многогранника. Прежде чем перейти к этой лемме, докажем вспомогательное утверждение.

Лемма 3. *Пусть дан многогранник M, заданный системой*

$$\begin{cases} f_i(x) \geqslant 0,\ i = 1, \ldots, m \\ \sum\limits_s x^{(s)} = 1 \end{cases},$$

для некоторых линейных функций f_i. Также пусть определено множество

$$I = \{i \,|\, f_i(v) = 0\}.$$

Тогда, чтобы являться вершиной многогранника M, для точки v необходимо и достаточно, чтобы система

$$\begin{cases} f_i(x) = 0,\ i \in I \\ f_i(x) > 0,\ i \notin I \\ \sum\limits_s x^{(s)} = 1 \end{cases} \tag{3.1}$$

имела единственное решение.

Доказательство.

Достаточность.

Предположим, что v не вершина, но при этом выполняются условия

$$f_i(x) \geqslant 0,\ i = 1 \ldots m.$$

В этом случае точка v принадлежит многограннику M. Так как она не является вершиной, то, по определению, она является серединой некоторого отрезка с концами u_1 и u_2, отличными от v. Это означает, что

$$v = \frac{u_1 + u_2}{2}.$$

В силу линейности f_i будет выполнено

$$\forall i \in I\ f_i(v) = \frac{1}{2}\left(f_i(u_1) + f_i(u_2)\right).$$

Но по определению множества I выполняется $f_i(v) = 0$, поэтому

$$\forall i \in I\ f_i(u_1) = -f_i(u_2).$$

При этом

$$\forall x \in M \ f_i(x) \geqslant 0 \text{ и } \sum_s x^s = 1.$$

Значит,

$$\forall i \in I \ f_i(u_1) = f_i(u_2) = 0.$$

Но тогда для любого $\alpha \in [0,1]$ точка $\alpha u_1 + (1-\alpha)u_2$ является решением системы (3.2). Противоречие с единственностью решения системы (3.2).

Необходимость.

Предположим, что решение системы (3.2) не единственно. Значит, множество точек задаваемых системой , является многогранником M', не вырожденным в точку. Все линейные функции, задающие грани M' (f_i при $i \notin I$), строго положительны в точке v, поэтому, по определению, точка v является внутренней точкой многогранника M'. Следовательно, существует отрезок $[u_1, u_2] \subset M' \subset M$ такой, что $v = \frac{1}{2}(u_1 + u_2)$. Значит, точка v не является вершиной M. □

Теперь мы можем сформулировать главную лемму, введя дополнительные обозначения

Определение 16. *Пусть $v \in \mathbb{R}^n$, тогда за $\operatorname{supp}(v)$ будет обозначаться множество позиций ненулевых элементов вектора v, а за $\overline{\operatorname{supp}(v)}$, соответственно, множество позиций нулевых элементов вектора v.*

Определение 17. *Пусть дана матрица $X \in \mathbb{R}^{n\times m}$, тогда за X_j будет обозначаться j-ый столбец матрицы X, а за $X\big[[i_1, \ldots, i_s], [j_1, \ldots, j_t]\big]$ будет обозначаться подматрица, состоящая из строк $i_1, \ldots, i_s$ и столбцов $j_1, \ldots, j_t$*

Лемма 4. *Пусть дана матрица$F \in \mathbb{R}^{n\times m}$ и её стохастическое разложение $F = \Phi\Theta$. Тогда, чтобы вершина Φ_j многогранника $\operatorname{conv}(\Phi)$ являлась вершиной многогранника $\operatorname{span}(\Phi) \cap \Delta_{n-1}$ необходимо и достаточно*

$$\operatorname{rank}\Big(\Phi\big[\overline{\operatorname{supp}(\Phi_k)}, [1,\ldots,k] \setminus [j]\big]\Big) = k-1.$$

Доказательство.
Для удобства можно рассматривать случай $j = k$ без ограничения общности доказательства.

Также обозначим набор утверждений, эквивалентность которых будет доказана далее:

(I) Φ_k является вершиной многогранника $\mathrm{span}(\Phi) \cap \Delta_{n-1}$

(II) $\exists!$ решение системы

$$\begin{cases} \sum\limits_{s=1}^{k} \Phi_{is} x^{(s)} = 0,\ i \in \overline{\mathrm{supp}(\Phi_k)} \\ \sum\limits_{s=1}^{k} \Phi_{is} x^{(s)} > 0,\ i \in \mathrm{supp}(\Phi_k) \\ \sum\limits_{s=1}^{k} x^{(s)} = 1 \end{cases} \tag{3.2}$$

(III) $\exists!$ решение системы уравнений

$$\sum_{s=1}^{k-1} \Phi_{is} x^{(s)} = 0,\ i \in \overline{\mathrm{supp}(\Phi_k)} \tag{3.3}$$

(IV) $\mathrm{rank}\Big(\Phi\big[\overline{\mathrm{supp}(\Phi_k)}, [1,\ldots,k] \setminus [k]\big]\Big) = k-1$

(III) $\Leftrightarrow$ (IV)

Система (3.3) всегда имеет нулевое решение. То, что это решение единственно, равносильно тому, что ядро отображения, задаваемого матрицей $\Phi\big[\overline{\mathrm{supp}(\Phi_k)}, [1,\ldots,k] \setminus [k]\big]$, нулевое, что равносильно (IV).

(III) $\Rightarrow$ (II)

При $i \in \overline{\mathrm{supp}(\Phi_k)}$ выполняется $\Phi_{ik} = 0$, откуда

$$\sum_{s=1}^{k} \Phi_{is} x^{(s)} = 0,\ i \in \overline{\mathrm{supp}(\Phi_k)} \Leftrightarrow \sum_{s=1}^{k-1} \Phi_{is} x^{(s)} = 0,\ i \in \overline{\mathrm{supp}(\Phi_k)}.$$

Таким образом, система (3.2) имеет больше условий чем система (3.3). Это означает, что количество решений системы (3.2) не превышается количество решений системы (3.3). В данном случае получается, что у системы (3.2) не более одного решения. Но вектор $(0,\ldots,0,1)$ всегда является решением системы (3.2). Значит, система (3.2) имеет единственное решение.

(II) $\Rightarrow$ (III)

Точка 0 (как вектор размерности $k-1$) всегда является решением системы (3.3). Пусть у системы (3.3) нашлось ещё одно решение $u \neq 0$. Тогда мы можем определить решение $\overline{u}$ системы (3.2), подобрав константу c в векторе

$\tilde{u} = (u^{(1)}, \ldots, u^{(k-1)}, c)$ и отнормировав его.

$$\overline{u} = \begin{cases} norm(\tilde{u}),\ if\ \sum\limits_s \tilde{u}^{(s)} \neq 0 \\ norm((\tilde{u}^{(1)}, \ldots, \tilde{u}^{(k-1)}, \tilde{u}^{(k)} + 1)),\ if\ \sum\limits_s \tilde{u}^{s} = 0 \end{cases},$$

где

$$norm(x)^{(i)} = \frac{x^{(i)}}{\sum\limits_s x^{(s)}},$$

$$c = \max(0, \tilde{c} + 1)$$

$$\tilde{c} = -\min_{w \in \operatorname{supp} \Phi_k} \frac{\sum\limits_{s=1}^{k-1} \Phi_{ws} u^{(s)}}{\Phi_{wk}}$$

Однако, у системы системы (3.2) есть ещё одно решение — вектор $(0, \ldots, 0, 1)$, и вектор $\overline{u}$ точно с ним не совпадает, так как содержит хотя бы одно ненулевое значение на первых $k-1$ позиции. Это противоречит единственности решения системы (3.2), значит, предположение было неверным и система (3.3) имела единственное нулевое решение.

(I) $\Leftrightarrow$ (II)

Следует из леммы 3. Заметим, что система

$$\begin{cases} \sum\limits_{s=1}^{k} \Phi_{is} x^{(s)} \geqslant 0 \\ \sum\limits_{s=1}^{k} x^{(s)} = 1 \end{cases}$$

задаёт в точности многогранник $\operatorname{span}(\Phi) \cap \Delta_n$ в пространстве $\operatorname{span}(\Phi)$, который соответствует многограннику M из леммы 3. Множеству I из леммы 3 соответствует $\overline{\operatorname{supp}(\Phi_k)}$. □

Теперь может быть доказана главная теорема главы о достаточных условиях единственности стохастического матричного разложения:

Теорема 13. *Пусть дано разложение* $F = \Phi\Theta$, $F \in \mathbb{R}^{n \times m}$, $\operatorname{rank} F = k$, $\Phi \in \mathbb{R}^{n \times k}$, $\Theta \in \mathbb{R}^{k \times m}$. *Пусть выполнены условия:*

– $\forall i \in \{1, \ldots, k\}\ \exists j : \Theta_{ij} = 1, \forall i' \neq j\ \Theta_{i'j} = 0$;

– $\forall j \;\; \operatorname{rank}\Big(\Phi\big[\overline{\operatorname{supp}(\Phi_j)}, [1,\ldots,k] \setminus [j]\big]\Big) = k-1.$

Тогда разложение $F = \Phi\Theta$ *единственно.*

Доказательство.

Условие

$$\forall i \in [1,\ldots,k] \; \exists j : \Theta_{ij} = 1, \forall i' \neq j \;\; \Theta_{i'j} = 0$$

означает, что существует k точек F таких, что они являются вершинами многогранника $\operatorname{conv}(\Phi)$. Откуда следует, что $\operatorname{conv}(F) = \operatorname{conv}(\Phi)$.

Из условия

$$\forall j \;\; \operatorname{rank}\Big(\Phi\big[\operatorname{supp}(\Phi_j), [1,\ldots,k] \setminus [j]\big]\Big) = k-1$$

следует, что каждая вершина $\operatorname{conv}(\Phi)$ является вершиной многогранника $\operatorname{span}(\Phi) \cap \Delta_{n-1}$ (Лемма 4). Далее применение Следствия 6 доказывает утверждение Теоремы. □

Условие на матрицу Φ легко обобщить на матрицу F, тем самым получив достаточное условие для единственности стохастического разложения.

Следствие 7. *Пусть дана стохастическая матрица* $F \in \mathbb{R}^{n\times m}, \operatorname{rank} F = k$. *Пусть также нашлось множество столбцов* $J = \{j_1,\ldots,j_k\}$ *для которых выполнено*

– $\forall j \in J$, $p = 1,\ldots,m$, *существует набор* $a_{jp} \geqslant 0$ *т. ч.* $\forall p \;\; \sum\limits_{j\in J} a_{jp} = 1$,

$F_p = \sum\limits_{j\in J} a_{jp} F_j.$

– $\forall j \in J \;\; \operatorname{rank}\Big(F\big[\overline{\operatorname{supp}(F_j)}, J \setminus [j]\big]\Big) = k-1,$

Тогда у F *существует единственное разложение* $F = \Phi\Theta$, *где*

$$\Phi = F[:, J],$$
$$\Theta[j,p] = a_{jp}.$$

Доказательство.

Возьмём в качестве матрицы Φ матрицу образованную столбцами $F_{j_1},\ldots,F_{j_k}$, а в качестве матрицы Θ матрицу образованную величинами a_{jp}. Первое условие обеспечивает выполнение первого условия Теоремы 13, а второе условие, соответственно второго условия Теоремы 13. Также из первого условия следует, что $F = \Phi\Theta$. □

Для сравнения Теорем 13 и 12 рассмотрим следующий пример:

$$F = \begin{pmatrix} 0 & \frac{1}{6} & \frac{2}{6} \\ 0 & \frac{2}{6} & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{2}{6} \\ \frac{2}{6} & 0 & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{6} & 0 \\ \frac{2}{6} & \frac{1}{6} & 0 \end{pmatrix}$$

Матрица F имеет единственное разложение ($F = FE$), при этом условия Теоремы 12 не выполнены, а условия следствия Теоремы 13 выполняются.

3.3 Эксперименты про проверке выполнения достаточных условий теоремы о единственности стохастического матричного разложения

С точки зрения тематического моделирования для условий Теоремы 13 возможна следующая интерпретация.

Условие 1 требует наличия в матрице Θ единичной подматрицы размера $k \times k$. Матрица Θ отвечает за распределение тем в документах. Поэтому фактически это условие требует наличия в тематической модели k унитематических документов, то есть таких, в которых есть одна тема с вероятностью появления 1, а вероятности остальных тем нулевые. Выполнение этого условия можно гарантировать, добавив в коллекцию k искусственно созданных унитематических документов, слова для которых подбираются, например, экспертами.

Условие 2 говорит о том, что для любого j произведение матриц

$$\Phi\left[\overline{\text{supp}(\Phi_j)}, [1, \ldots, k] \setminus [j]\right] \text{ и } \Theta\left[[1, \ldots, k] \setminus [j], :\right]$$

является неотрицательным матричным разложением полного ранга для матрицы

$$F\left[\overline{\text{supp}(\Phi_j)}, :\right].$$

С точки зрения тематического моделирования это означает, что если для любой темы t из матрицы слова-документы F ранга T убрать все слова, встречающиеся в t-ой теме, то на получившей матрице слова-документы можно построить невырожденную тематическую модель на $T - 1$ теме.

Для проверки выполнения второго условия на реальной текстовой коллекции был проведен эксперимент.

3.3.1 Описание эксперимента

Эксперимент проводился на лемматизированной коллекции 20Newsgroups [48]. Для проверки условий теоремы использовалась исходная матрица встречаемости слова-документы, а для оценки устойчивости ещё искусственная матрица, полученная по коллекции следующим способом: было найдено решение (Φ, Θ) оптимизационной задачи, а затем использовалась матрица $\Phi\Theta$. Особенностью этой матрицы являлось то, что она обеспечивала глобальное и единственное (так как на нём выполнялись условия Теоремы 13) решение оптимизационной задачи.

При проверке выполнения условий теоремы проверялось только второе условие, так как выполнение первого обеспечивалось за счёт обогащения коллекции псевдодокументами, по одному для каждой темы. Псевдодокумент для фиксированной темы строился следующим образом. В него вручную добавлялись (некоторые) слова данной темы, после чего в матрице Θ для этого документа назначались нулевые значения всем темам кроме исходной. Добавление таких псевдодокументов по одному для каждой темы обеспечивает выполнение первого условия теоремы.

Тематическая модель строилась алгоритмом оптимизации ARTM ([51], [52]) с регуляризатором разреживания

$$R(\Phi) = \alpha \sum_t \sum_w \ln \varphi_{wt}$$

с коэффициентом регуляризации α. В использованной реализации алгоритма ARTM по умолчанию матрица Φ не может содержать нулей. Регуляризатор разреживания позволяет добиться зануления малых значений в этой матрице и контролировать количество нулей.

При оценке устойчивости сравнивались 4 различных схемы:

1. Раскладывается исходная матрица слова-документы, никаких ограничений на начальную инициализацию. (далее обозначается PLSA).

2. Раскладывается исходная матрица слова-документы, первый запуск определял матрицы Φ и Θ, фиксировалось множество нулевых элементов обоих матриц, и при последующих начальных инициализациях эти элементы оставались нулевыми. Так как в алгоритме PLSA если какой-то элемент Φ или Θ равен нулю, то он будет оставаться равен нулю на последующих итерациях, то за счёт инициализации мы можем обеспечивать ограничение на множестве нулевых элементов в полученном решении оптимизационной задачи. (далее Init PLSA).
3. Раскладывается искусственная матрица слова-документы, никаких ограничений на начальную инициализацию. (далее synPLSA).
4. Раскладывается искусственная матрица слова-документы, начальная инициализация аналогично пункту 2. (далее Init synPLSA).

Для каждой схемы проверялись 100 различных начальных случайных инициализаций, а затем оценивались свойства полученных наборов матриц Φ, и также визуализировалось их расположение при помощи алгоритма tSNE [53].

Для удобства обозначим

$$\Phi[\overline{supp(\Phi_t)}, \{1, \ldots, T\} \setminus \{t\}]$$

за матрицу U_t. Для проверки второго условия теоремы требовалось находить для каждой темы t величину $\operatorname{rank} U_t$ Эффективный вычислительный способ сделать это — нахождение минимального сингулярного значения матрицы U. Эта величина в дальнейшем будет называться uniqueness measure темы t. Геометрически она означает насколько не плоским является многогранный угол при вершине темы t в многограннике, натянутом на вектора тем Φ_t. Если эта величина больше 0, то можно утверждать, что матрица U_t имеет полный ранг.

Для большей интерпретируемости uniqueness measure стоит отнормировать на минимальное сингулярное значение матрицы

$$\Phi[\{1, \ldots, W\}, \{1, \ldots, T\} \setminus \{t\}].$$

Полученная величина будет лежать в промежутке от 0 до 1, причём 1 будет достигаться тогда и только тогда, когда матрица

$$\Phi[supp(\Phi_t), \{1, \ldots, T\} \setminus \{t\}]$$

нулевая, что означает полное отделение слов темы. Далее эта величина будет называться normalized uniqueness measure.

Для оценки устойчивости решения оптимизационной задачи использовались две группы метрик. Как было сказано, имеется набор матриц Φ. Далее для этого набора оценивался разброс, как среднее попарное расстояние между этими матрицами, также в схемах 2-4 оценивалось смещение, как среднее расстояние до истинной матрицы Φ, при помощи которой генерировалась коллекция.

Расстояние между матрицами рассчитывалось следующим образом:

$$\rho(\Phi^1, \Phi^2) = \min_S \left(\sum_i \rho^0(\Phi^1_i, (\Phi^2 S)_i) \right),$$

где S — перестановочные матрицы размера $T \times T$, а ρ^0 — одна из следующих функций расстояния между распределениями:

$$L_1(p, q) = \sum_k |p_k - q_k|,$$

$$\text{sMAPE}(p, q) = \frac{1}{n} \sum_k \frac{|p_k - q_k|}{|p_k| + |q_k|},$$

$$\text{KL}(p, q) = \sum_k p_k \log \frac{p_k}{q_k},$$

$$\text{KL}_2(p, q) = \text{KL}(p, z) + \text{KL}(q, z), \text{ где } z = \frac{p + q}{2}.$$

3.3.2 Результаты

Как показывает Рисунок 3.1 даже незначительного разреживания ($\alpha = -10^{-25}$) достаточно, чтобы обеспечить единственность разложения. Однако, если запустить исходный алгоритм PLSA без какого либо разреживания, то условия теоремы не будут выполняться, поскольку в данном алгоритме не может происходить зануление элементов матриц. Тем не менее, это означает, что та точка, к которой стремится решение PLSA, удовлетворяет условиям теоремы.

На Рисунке 3.2 можно увидеть, что решения не имеют какой-то общей структуры. Так как tSNE учитывает локальные особенности точек, то по этим изображениям нельзя сделать вывод о силе разброса точек, можно лишь

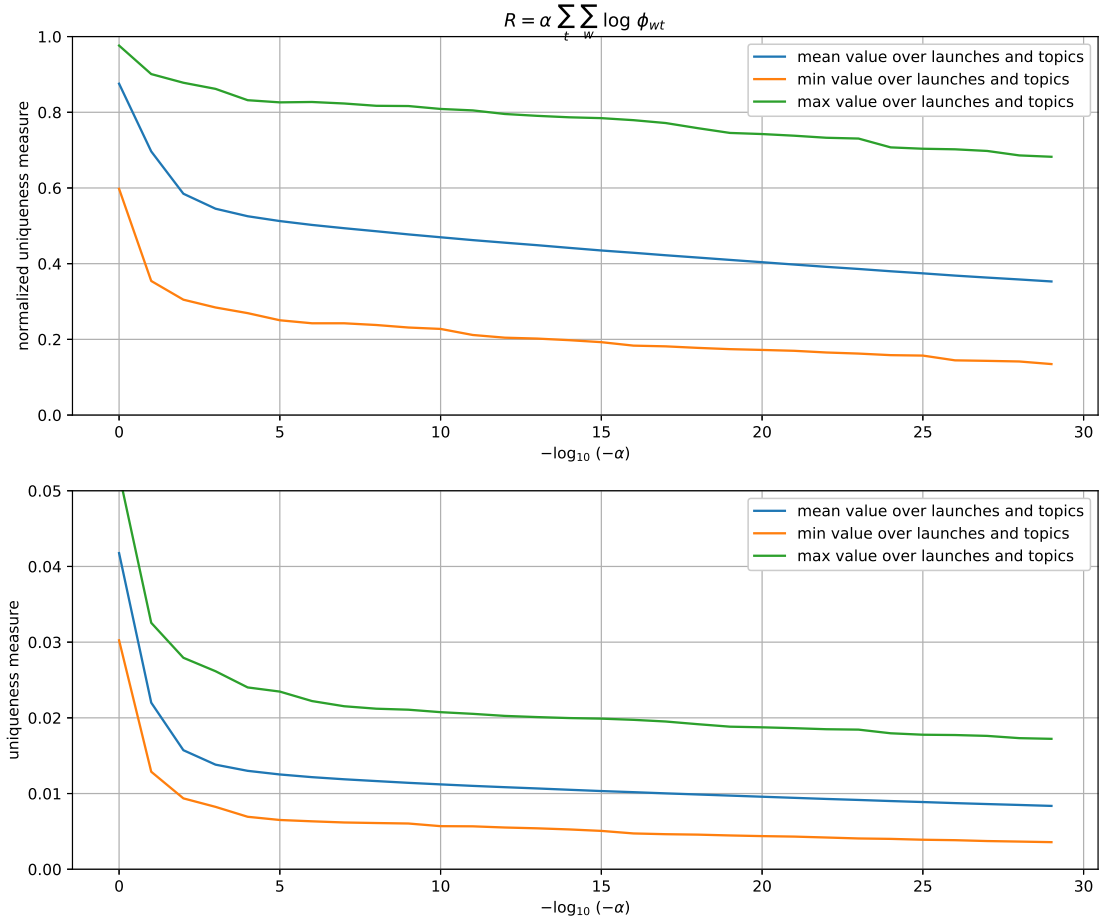

Рисунок 3.1 — Зависимость uniqueness measure и normalized uniqueness measure от коэффициента разреживания α.

Value	PLSA	Init PLSA	synPLSA	Init synPLSA
Variance L_1	0.6733 ± 0.0027	0.0264 ± 0.0002	0.1373 ± 0.0128	0.0000 ± 0.0001
Variance sMAPE	1.2952 ± 0.0025	0.0696 ± 0.0003	1.2473 ± 0.0137	0.0003 ± 0.0001
Variance KL	∞	∞	0.3310 ± 0.0486	$(0.0000 \pm 0.0001$
Variance KL_2	0.2845 ± 0.0015	0.0022 ± 0.0001	0.0442 ± 0.0045	0.0000 ± 0.0001
Bias L_1	—	0.0253 ± 0.0006	0.0959 ± 0.0259	0.0000 ± 0.0001
Bias sMAPE	—	0.0656 ± 0.0010	1.3867 ± 0.0129	0.0030 ± 0.0001
Bias KL	—	∞	0.0553 ± 0.0159	0.0000 ± 0.0001
Bias KL_2	—	0.0020 ± 0.0001	0.0331 ± 0.0095	0.0000 ± 0.0001

Таблица 2 — Средние и доверительные интервалы для метрик устойчивости

проверить наличие кластеров. В случае Init PLSA вы видим явные кластера различных решений, что означает, что, несмотря на выполнение условий теоремы, неединственность обеспечивается за счёт неединственного решения оптимиза-

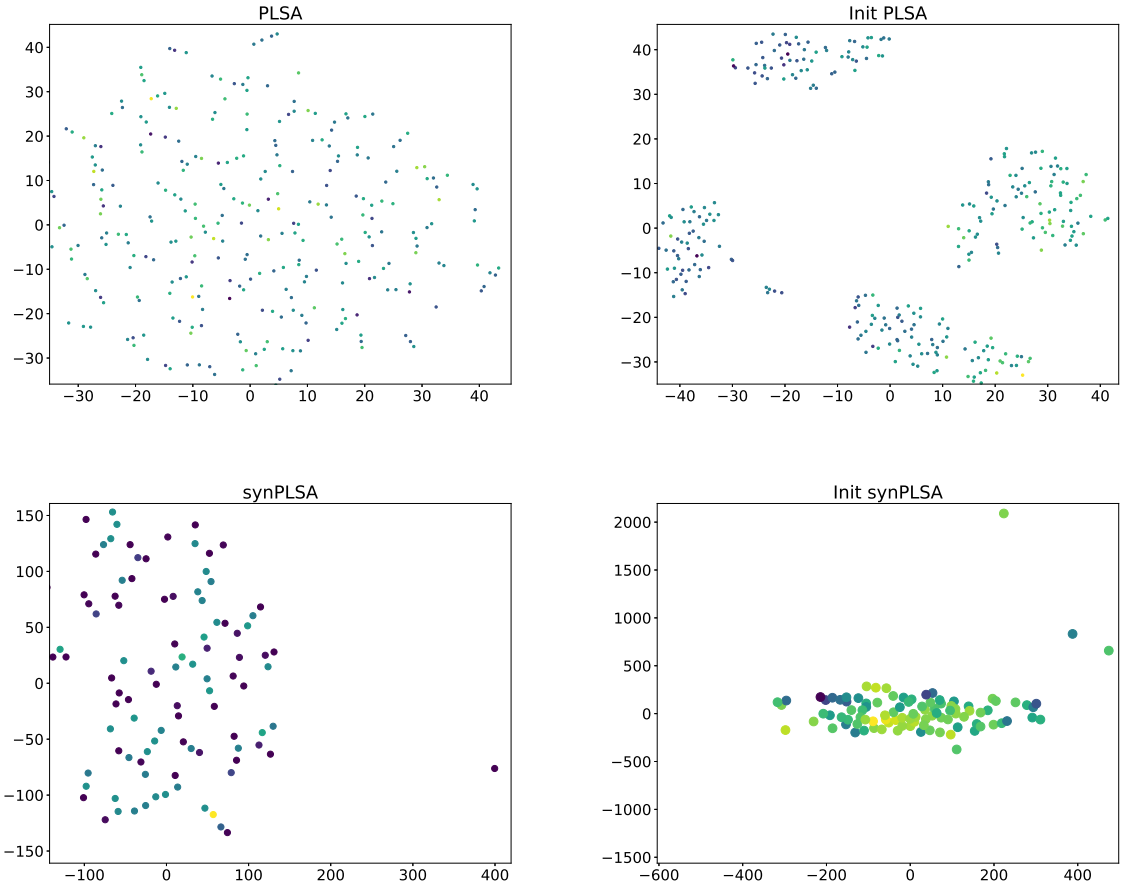

Рисунок 3.2 — Визуализация устойчивости при помощи алгоритма tSNE.

ционной задачи, причём эти решения могут существенно отличаться друг от друга.

В Таблице 2 приведены средние с доверительными интервалами для исследуемых метрик разброса и смещения. В случаях PLSA и synPLSA значения Variance очень большие, так как без каких-либо ограничений на множество нулевых элементов матриц решений оптимизационной задачи крайне много и все эти решения различны.

Однако, если зафиксировать множество нулевых элементов, то уже даже в случае исходной матрицы (Init PLSA) можно наблюдать сильное уменьшение разброса и восстановление исходной матрицы Φ с достаточно хорошей точностью. Тем не менее, решение оптимизационной задачи может быть нединственно даже при фиксации нулевых элементов матриц. Поэтому значение Variance не опускается до нуля в случае Init PLSA. Но когда решение оптимизационной задачи единственно при фиксации множества нулевых элементов (Init synPLSA), то можно наблюдать устойчивое нахождение исходной матрицы Φ.

3.4 Заключение главы

Была описана проблема неединственности решения в задаче тематического моделирования, которая декомпозируется на неоднозначность выбора локального экстремума $\tilde{F} = \Phi\Theta$ оптимизируемого функционала и неединственность разложения $\tilde{F}$ на Φ и Θ.

Был сформулирован ранее неизвестный результат (Теорема 13), дающий достаточные условия для единственности решения задачи стохастического матричного разложения. Также были реализованы эксперименты (раздел 3.3), в которых подтвердилось выполнение условий Теоремы 13 на реальной текстовой коллекции. Таким образом, было показано, что неединственность решения в задаче тематического моделирования возникает в основном из-за неоднозначности выбора локального экстремума $\tilde{F}$.

Глава 4. Разреживание тематических моделей

Теорема 7 о достаточных условиях сходимости выделяет важность множества нулей, которое фиксируется с некоторой итерации в EM-алгоритме ARTM. Дальнейшая оптимизация ведётся при ограничение на это множество, соответственно, оно во многом определяет качество полученного решения.

Теорема 13 предлагает достаточные условия для единственности стохастического матричного разложения полученного решения оптимизационной задачи ARTM. Чем выше уровень разреженности матрицы Φ, тем выше вероятность выполнения условий теоремы.

Таким образом, оказывается, что разреженность очень важна для повышения качества решения задачи тематического моделирования. А, значит, появляется проблема увеличения разреженности для произвольных матрицы Φ и Θ, при условии повышения или незначительного понижения правдоподобия (1.1).

В этой главе для решения поставленной проблемы предлагается применить метод аналогичный методу Optimal Brain Damage (OBD [54]) для обучения нейросетей.

4.1 Описание метода

Решаемую задачу можно сформулировать следующим способом. Обозначим за L функцию логарифма правдоподобия:

$$L(\Phi, \Theta) = \sum_{d,w} n_{dw} \log \sum_{t} \varphi_{wt}\theta_{td}.$$

Имеется некоторая пара матриц Φ и Θ, требуется сделать их более разреженным, сохранив структуру и несильно уменьшив функционал L.

Основная идея метода заключается в том, что стоит занулять те значения φ_{wt} и θ_{td}, при зануленении которых значение L изменится наименьшим способом.

Утверждение 6. *При занулении значения φ_{wt} эффект уменьшения значения L равен*

$$\Delta_1 L = -n_{wt}\left(1 + O\left(\max_d p_{tdw}\right)\right).$$

Доказательство.

$$\Delta_1 L = \sum_d n_{dw}\left(\log\sum_{s\neq t}\varphi_{ws}\theta_{sd} - \log\sum_s \varphi_{ws}\theta_{sd}\right) =$$

$$= \sum_d n_{dw}\log\frac{\sum\limits_{s\neq t}\varphi_{ws}\theta_{sd}}{\sum_s \varphi_{ws}\theta_{sd}} = \sum_d n_{dw}\log\left(1 - \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}\right) =$$

$$= \sum_d n_{dw}\log(1 - p_{tdw}) = \sum_d n_{dw}\left(-p_{tdw} - O(p_{tdw}^2)\right) =$$

$$= -\sum_d n_{dw}p_{tdw}\left(1 + O(p_{tdw})\right) = -n_{wt}\left(1 + O\left(\max_d p_{tdw}\right)\right).$$

□

Таким образом, если мы хотим занулять параметры модели для разреживания, на первый взгляд кажется, что нужно занулять параметры с наименьшим значением n_{wt}. Однако, здесь опускается один важный момент: когда мы разреживаем модель и зануляем какой-то φ_{wt}, вероятность всех остальных слов в теме увеличивается, а это увеличивает L.

Утверждение 7. *При занулении значении φ_{wt} эффект увеличения значения L равен*

$$\Delta_2 L = \frac{\varphi_{wt}n_t}{1-\varphi_{wt}}\left(1 - \frac{n_{wt}}{n_t}\right)\left(1 + O\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\frac{n_{wt}}{n_t}\max_{d,u} p_{tdu}\right)\right).$$

Доказательство.

$$\Delta_2 L = \sum_{d,u\neq w} n_{du}\left(\log\left(\sum_{s\neq t}\varphi_{us}\theta_{sd} + \frac{\varphi_{ut}}{1-\varphi_{wt}}\theta_{td}\right) - \log\sum_s \varphi_{us}\theta_{sd}\right) =$$

$$= \sum_{d,u\neq w} n_{du}\left(\log\left(\sum_s \varphi_{us}\theta_{sd} + \frac{\varphi_{wt}}{1-\varphi_{wt}}\varphi_{ut}\theta_{td}\right) - \log\sum_s \varphi_{us}\theta_{sd}\right) =$$

$$= \sum_{d,u\neq w} n_{du}\log\left(1 + \frac{\varphi_{wt}}{1-\varphi_{wt}}\frac{\varphi_{ut}\theta_{td}}{\sum_s \varphi_{us}\theta_{sd}}\right) = \sum_{d,u\neq w} n_{du}\log\left(1 + \frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu}\right) =$$

$$= \sum_{d,u\neq w} n_{du}\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu} + O\left(\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu}\right)^2\right)\right) =$$

$$= \frac{\varphi_{wt}}{1-\varphi_{wt}}\sum_{d,u\neq w} n_{du}p_{tdu}\left(1 + O\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu}\right)\right) =$$

$$= \frac{\varphi_{wt}}{1-\varphi_{wt}}\left(n_t - n_{wt} + O\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\sum_{d,u\neq w} n_{du}p_{tdu}^2\right)\right) =$$

$$= \frac{\varphi_{wt}n_t}{1-\varphi_{wt}}\left(1 - \frac{n_{wt}}{n_t} + O\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\frac{n_{wt}}{n_t}\max_{d,u} p_{tdu}\right)\right) =$$

$$= \frac{\varphi_{wt}n_t}{1-\varphi_{wt}}\left(1 - \frac{n_{wt}}{n_t}\right)\left(1 + O\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\frac{n_{wt}}{n_t}\max_{d,u} p_{tdu}\right)\right).$$

□

Если матрицы Φ и Θ брать ближе к последним итерациям алгоритма, то $\varphi_{wt} \approx \frac{n_{wt}}{n_t}$ откуда следует, что $\Delta_2 L \approx n_{wt}$. Таким образом, суммарно в первом приближении $\Delta_1 L + \Delta_2 L \approx 0$, то есть L не изменится.

Первый способ решения этой проблемы — точно рассчитать ΔL.

Теорема 14. *Изменение значения L при зануленении значения φ_{wt} составляет*

$$\Delta L = \sum_d n_{dw}\log(1 - p_{tdw}) + \sum_{d,u\neq w} n_{du}\log\left(1 + \frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu}\right). \tag{4.1}$$

Изменение значения L при занулении значения θ_{td} составляет

$$\Delta L = \sum_w n_{dw}\log\left(1 - p_{tdw}\right) - n_d\log(1 - \theta_{td}). \tag{4.2}$$

Доказательство.

Зануление значения φ_{wt}.

В доказательстве Утверждения 6 было показано, что

$$\Delta_1 L = \sum_d n_{dw}\log(1 - p_{tdw}).$$

В доказательстве Утверждения 7

$$\Delta_2 L = \sum_{d,u\neq w} n_{du}\log\left(1 + \frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu}\right).$$

Сложив $\Delta_1 L$ и $\Delta_2 L$, получим (4.1).

Зануление значения θ_{td}.
При занулении θ_{td} изменение L составит:

$$\Delta L = \sum_w n_{dw} \left(\log \left(\frac{1}{1-\theta_{td}} \sum_{s \neq t} \varphi_{ws}\theta_{sd} \right) - \log \sum_s \varphi_{ws}\theta_{sd} \right) =$$

$$= \sum_w n_{dw} \log \left(\frac{1}{1-\theta_{td}} \left(1 - \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}} \right) \right) = \sum_w n_{dw} \log \frac{1-p_{tdw}}{1-\theta_{td}} =$$

$$= \sum_w n_{dw} \log (1-p_{tdw}) - n_d \log(1-\theta_{td}).$$

□

Однако, подобные вычисления могут быть, во-первых, вычислительно долгими из-за взятия логарифмов и существенно замедлять алгоритм, а во-вторых, влиять на асимптотику вычисления. Например, выражение $\Delta_2 L$ при зануленииии φ_{wt} вычисляется за $O(|N||W|)$, где $|N|$ — суммарная длина коллекции, что явно дольше времени работы E-шага и будет замедлять алгоритм. Для оценки ΔL можно выполнить разложение до второго члена в ряде Тейлора слагаемых (по аналогии с OBD) и получить примерные оценки.

Теорема 15. *Изменение значения L при зануленииии значения φ_{wt} после аппроксимации составляет*

$$\Delta L = \frac{n_t \varphi_{wt} - n_{wt}}{1-\varphi_{wt}} - \frac{1}{2} \left(\frac{\varphi_{wt}}{1-\varphi_{wt}} \right)^2 \sum_{d, u \neq w} n_{du} p_{tdu}^2 + \frac{1}{2} \sum_d n_{dw} p_{tdw}^2 + O\left(\sum_d n_{dw} p_{tdw}^3 \right).$$

Изменение значения L при зануленииии значения θ_{td} после аппроксимации составляет

$$\Delta L = (n_d \theta_{td} - n_{td}) + \frac{1}{2} \sum_w n_{dw} p_{tdw}^2 - \frac{1}{2} n_d \theta_{td}^2 + O\left(\sum_w n_{dw} p_{tdw}^3 \right).$$

Доказательство.
Зануление φ_{wt}.
Оценим первое слагаемое (4.1):

$$\sum_d n_{dw} \log(1-p_{tdw}) = \sum_d n_{dw} \left(-p_{tdw} + \frac{1}{2} p_{tdw}^2 + O(p_{tdw}^3) \right) =$$

$$= -n_{wt} + \frac{1}{2}\sum_d n_{dw}p_{tdw}^2 + O\left(\sum_d n_{dw}p_{tdw}^3\right).$$

Оценим второе слагаемое (4.1):

$$\sum_{d,u\neq w} n_{du}\log\left(1+\frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu}\right) =$$

$$= \sum_{d,u\neq w} n_{du}\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu} - \frac{1}{2}\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\right)^2 p_{tdu}^2 + O\left(\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}p_{tdu}\right)^3\right)\right) =$$

$$= \frac{\varphi_{wt}}{1-\varphi_{wt}}(n_t - n_{wt}) - \frac{1}{2}\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\right)^2 \sum_{d,u\neq w} n_{du}p_{tdu}^2 + \left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\right)^3 O\left(\sum_{d,u\neq w} n_{du}p_{tdu}^3\right).$$

Объединяя, получим

$$\Delta L = -n_{wt} + \frac{1}{2}\sum_d n_{dw}p_{tdw}^2 + O\left(\sum_d n_{dw}p_{tdw}^3\right) + \frac{\varphi_{wt}}{1-\varphi_{wt}}(n_t - n_{wt}) -$$

$$-\frac{1}{2}\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\right)^2 \sum_{d,u\neq w} n_{du}p_{tdu}^2 + \left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\right)^3 O\left(\sum_{d,u\neq w} n_{du}p_{tdu}^3\right) =$$

$$= \frac{n_t\varphi_{wt} - n_{wt}}{1-\varphi_{wt}} - \frac{1}{2}\left(\frac{\varphi_{wt}}{1-\varphi_{wt}}\right)^2 \sum_{d,u\neq w} n_{du}p_{tdu}^2 + \frac{1}{2}\sum_d n_{dw}p_{tdw}^2 + O\left(\sum_d n_{dw}p_{tdw}^3\right).$$

Зануление θ_{td}.

Оценим первое слагаемое (4.2):

$$\sum_w n_{dw}\log(1-p_{tdw}) = \sum_w n_{dw}\left(-p_{tdw} + \frac{1}{2}p_{tdw}^2 + O(p_{tdw}^3)\right) =$$

$$= -n_{td} + \frac{1}{2}\sum_w n_{dw}p_{tdw}^2 + O\left(\sum_w n_{dw}p_{tdw}^3\right).$$

Оценим второе слагаемое (4.2):

$$-n_d\log(1-\theta_{td}) = n_d\theta_{td} - \frac{1}{2}n_d\theta_{td}^2 + O(n_d\theta_{td}^3).$$

Объединяя получим

$$\Delta L = -n_{td} + \frac{1}{2}\sum_w n_{dw}p_{tdw}^2 + O\left(\sum_w n_{dw}p_{tdw}^3\right) + n_d\theta_{td} - \frac{1}{2}n_d\theta_{td}^2 + O(n_d\theta_{td}^3) =$$

$$= (n_d\theta_{td} - n_{td}) + \frac{1}{2}\sum_w n_{dw}p_{tdw}^2 - \frac{1}{2}n_d\theta_{td}^2 + O\left(\sum_w n_{dw}p_{tdw}^3\right).$$

□

Эти формулы являются приближением формул (4.1) и (4.2), они могут быть использованы, когда производительность критична. С алгоритмической точки зрения сложность представляет только выражение $\Delta_2 L$ при занулении φ_{wt}, поскольку не вычисляется за асимптотику E-шага $O(|N|)$ и потенциально замедляется асимптотику алгоритма. Теорема 15 предлагает приближение

$$\Delta_2 L = \alpha_{wt}(n_t - n_{wt}) - \frac{1}{2}\alpha_{wt}^2 \sum_{d,u\neq w} n_{du}p_{tdu}^2 + \alpha_{wt}^3 O\left(\sum_{d,u\neq w} n_{du}p_{tdu}^3\right),$$

где $\alpha_{wt} = \dfrac{\varphi_{wt}}{1-\varphi_{wt}}$. Это приближение вычисляется за $O(|N|)$, остальные слагаемые Теоремы 14 тоже вычисляются за эту асимптотику. Таким образом, используя приближения только для $\Delta_2 L$, полученные значения будут наиболее точны, а асимптотика алгоритма не будет отличаться от асимптотики E-шага, не меняя асимптотику алгоритма.

4.2 Описание экспериментов по разреживанию моделей

Есть несколько вариантов использования оценки изменения ΔL:

1. Разреживание имеющейся тематической модели.
2. Разреживание в ходе итераций EM-алгоритма ARTM.
3. Определение структуры разреженности.

Отличие первого и второго пункта состоит в том, что для готовой тематической модели будет выполнено $\varphi_{wt} \approx \frac{n_{wt}}{n_t}$ и $\theta_{td} \approx \frac{n_{td}}{n_d}$, поэтому точность приближения ΔL будет более важна. Для первого пункта в качестве начального приближения использовались матрицы Φ и Θ, полученные алгоритмом PLSA за 100 итераций. Для второго пункта в качестве начального приближения использовались матрицы Φ и Θ, полученные алгоритмом PLSA за 1 итерацию. Одна итерация была необходима, чтобы не применять OBD к случайным матрицам, которые являлись начальной инициализацией для PLSA.

Обозначим за γ_{wt} изменение $-\Delta L$ при зануления φ_{wt}. Под структурой разреженности понимается отображение всех ячеек матрицы Φ в осях n_{wt}, γ_{wt}. Это изображение показывает зависимость изменения логарифма правдоподобия от популярности слова в теме, а также на нём можно увидеть распределение термов по степени их влияния на оптимизируемый функционал.

В качестве текстовой коллекции для оценки поведения использовалась «20 NewsGroups».

Сравнивались три алгоритма:

1. **sparse LDA**. Разреживащий LDA, реализованный посредством регуляризатора разреживания. Зануление элемента φ_{wt} происходит, если $n_{wt} \leqslant 1$.
2. **OBD ARTM limited**. Стандартный алгоритм ARTM, дополнительно на каждой итерации рассчитывающий γ_{wt} по формулам Теорем 14 и 15. Зануление элемента φ_{wt} происходит, если $\gamma_{wt} \leqslant 1$.
3. **OBD ARTM**. Аналогично OBD ARTM limited, но дополнительно на каждой итерации зануляется 0.5% элементов с наибольшим значением γ_{wt}.

Полученные тематические модели оценивались по двум метрики:

Разреженность матрицы Φ. Разреженность определяется как доля нулей в матрице. Как отмечалось в начале главы, мы хотим значительно повысить разреженность моделей, так как это положительно сказывается на теоретических свойствах полученных решений. Также увеличение разреженности положительно влияет на интерпретируемость полученной модели и уменьшает переобучение.

Перплексия. Перплексия это стандартная метрика для вероятностных тематических моделей, она определяется по логарифму правдоподобия по следующей формуле:

$$perplexity(\Phi, \Theta) = \exp\left(-\frac{L(\Phi, \Theta)}{n}\right), \tag{4.3}$$

где $n = \sum_{d,w} n_{dw}$ — суммарная длина коллекции. Поскольку задача разреживания ставится как увеличение разреженности при неуменьшении L, то по перплексии определяется степень этого неуменьшения.

4.3 Результаты экспериментов по разреживанию моделей

$\|T\|$	Метрика	Алгоритм	До	После	Увеличение, %
10	Разреженность	sparse LDA	0.187	0.755	303
10	Разреженность	OBD ARTM	0.187	0.751	301
10	Разреженность	OBD ARTM limited	0.187	0.75	301
10	Перплексия	sparse LDA	2034.7	2374.0	16
10	Перплексия	OBD ARTM	2034.7	2064.3	1
10	Перплексия	OBD ARTM limited	2034.7	2059.9	1
25	Разреженность	sparse LDA	0.32	0.866	170
25	Разреженность	OBD ARTM	0.32	0.861	169
25	Разреженность	OBD ARTM limited	0.32	0.86	168
25	Перплексия	sparse LDA	1518.1	2121.5	39
25	Перплексия	OBD ARTM	1518.1	1553.9	2
25	Перплексия	OBD ARTM limited	1518.1	1549.8	2

Таблица 3 — Разреженность и перплексия после 1 итерации разреживания разными методами

$\|T\|$	Метрика	Алгоритм	До	После	Увеличение, %
10	Разреженность	sparse LDA	0.187	0.836	347
10	Разреженность	OBD ARTM	0.187	0.808	332
10	Разреженность	OBD ARTM limited	0.187	0.764	308
10	Перплексия	sparse LDA	2034.7	2214.7	8
10	Перплексия	OBD ARTM	2034.7	2092.7	2
10	Перплексия	OBD ARTM limited	2034.7	2037.7	0
25	Разреженность	sparse LDA	0.32	0.923	188
25	Разреженность	OBD ARTM	0.32	0.899	180
25	Разреженность	OBD ARTM limited	0.32	0.871	172
25	Перплексия	sparse LDA	1518.1	1754.7	15
25	Перплексия	OBD ARTM	1518.1	1580.2	4
25	Перплексия	OBD ARTM limited	1518.1	1521.5	0

Таблица 4 — Разреженность и перплексия после 100 итерации разреживания разными методами

Таблица 3 показывает, что зануление элементов матрицы Φ по величинам n_{wt} и по величинам γ_{wt} дают примерно одинаковое увеличение разреженности модели при разном количестве тем $|T|$. При этом зануление по γ_{wt} увеличивает перплексию в незначительной мере, в отличие от зануления по n_{wt}.

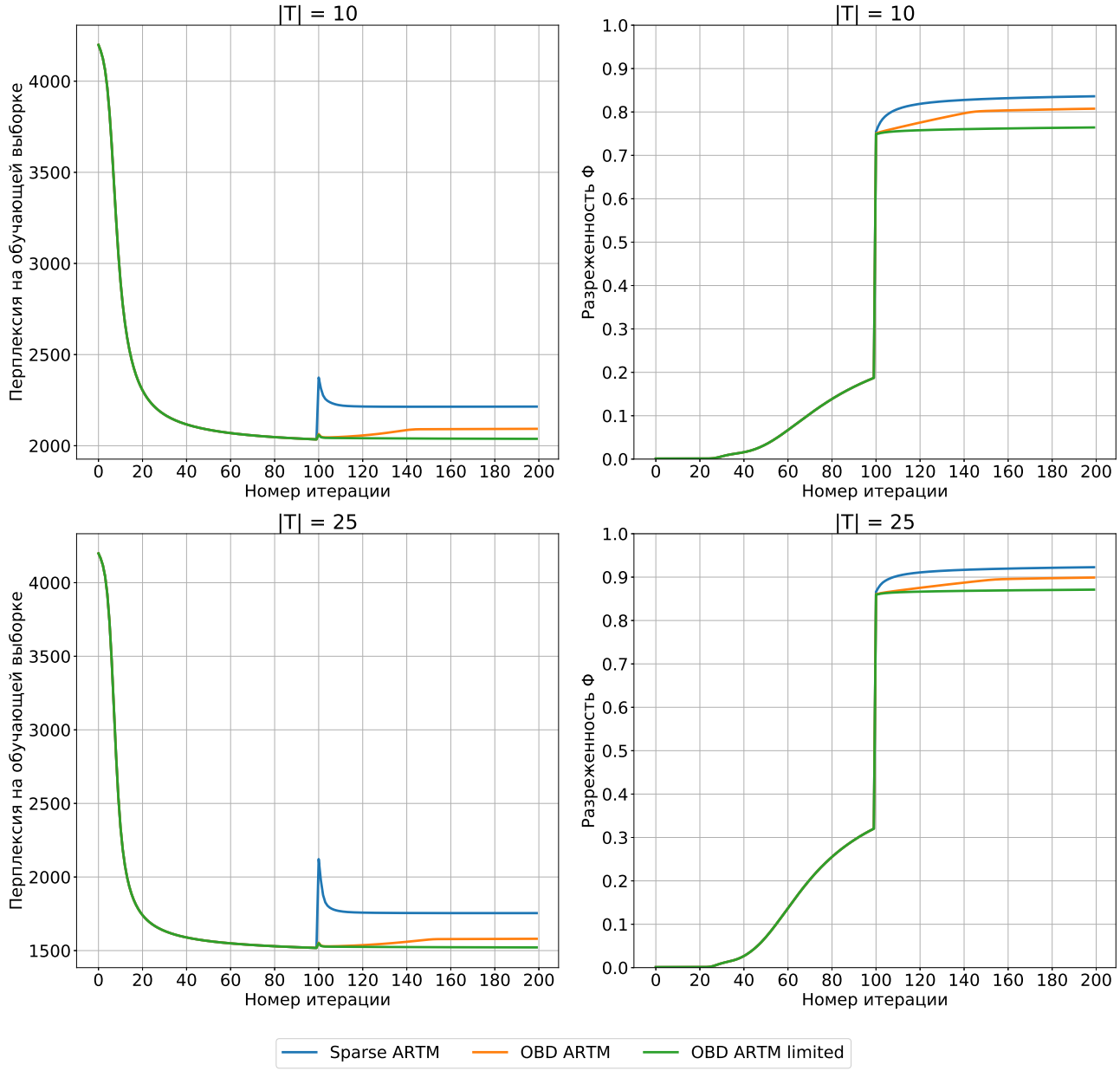

Рисунок 4.1 — Изменения метрик на итерациях для разреживающего LDA и двух версий алгоритма OBD ARTM.

Таблица 4 показывает, что с повторениями итераций LDA позволяет ещё дополнительно увеличить разреженность матриц и уменьшить перплексию. Это объясняется тем, что разреживающий LDA постепенно на итерациях подстраивается под данные и нивелирует сильную просадку перплексии на первых итерациях. Однако, OBD продолжает показывать слабое увеличение перплексии в отличие от LDA.

На Рисунке 4.1 более подробно показаны изменения перплексии и разреженности на итерациях. Первые 100 итераций графики совпадают, так как использовался общий алгоритм PLSA для получения матриц для разрежива-

ния. на 100-ой итерации мы видим резкий скачок метрик связанный с первой итерацией зануления. После чего алгоритмы адаптируются к произошедшим изменениям на продолжении последующих 100 итераций разрежения. Алгоритм OBD ARTM за счёт дополнительного зануления элементов по сравнению с OBD ARTM limited показывает чуть большее значение разреженности, но и чуть большее значение перплексии, однако, разница не значительна. Также на графике видно, что LDA требуется порядка 5 дополнительных итераций, чтобы уменьшить потерю в перплексии на первой итерации. В то же время OBD ARTM limited без дополнительных итераций даёт приемлемое значение перплексии.

На Рисунках 4.2 и 4.3 изображены изменения структуры разреженности в логарифмических шкалах на итерациях OBD ARTM.

Самый первый график показывает структуру разреженности для случайной инициализации матриц. При больших значениях n_{wt} выполняется $n_{wt} \approx \gamma_{wt}$, при малых же значениях такой корреляции не прослеживается, однако, $\gamma_{wt} \leqslant n_{wt}$. Это означает, что есть много пар тема-слово, таких что слово достаточно часто встречается в теме, но при занулении этого элемента, правдоподобие сильно не изменяется.

Графики с 20 по 100 итерацию показывают структуру разреженности для алгоритма PLSA. Мы видим облако точек вокруг прямой $n_{wt} = \gamma_{wt}$, причём облако значительно шире чем у случайных матриц. Это говорит о том, что корреляция между n_{wt} и γ_{wt} присутствует, но они явно не совпадают. Также на этих графиках видно большое облако точек при малых значениях n_{wt} и γ_{wt}. Это основные кандидаты на зануление, так как их эффект на правдоподобие очень мал. Видно, что при малых значениях n_{wt} выполняется $n_{wt} \geqslant \gamma_{wt}$. Это объясняется равенством (4.1), при малых n_{wt} первое слагаемое примерно равно $-n_{wt}$, а второе слагаемое положительно, откуда следует, что $-\Delta L \leqslant n_{wt}$.

На графиках с 120 по 220 итерацию видно, что произошло зануление элементов, слабо влияющих на правдоподобие. Чем больше занулений происходит, тем большая часть нижних точек отсекается. Можно наблюдать горизонтальную линию, выше которой находится почти всё облако точек, она соответствует порогу зануления на итерации. На графике 120-ой итерации он соответствует $\gamma_{wt} \leqslant 1$, на последующих итерациях он определяется процентилью ненулевых значений γ_{wt}, которая постепенно растёт.

Также образуется ступеньки на определённых значениях γ. Это объясняется тем, что при реализации алгоритма ARTM часто используется технический трюк — при вычислении логарифма правдоподобия L по формуле (1.1) к выражению под логарифмов добавляется небольшое значение ε чтобы избежать $L = -\infty$ на итерациях. Соответственно, если при занулении φ_{wt} зануляется выражение под логарифмов, то $\Delta L = n_w \log \varepsilon$. Поэтому ступеньки, которые мы видим на графике соответствуют значениям $\log n_w + \log\log \frac{1}{\varepsilon}$ при разных n_w от 1 и т.д. При реализация использовалось $\varepsilon = 10^{-20}$, а, значит, первая ступенька должна соответствовать $\log\log 10^2 0 \approx 3.83$, вторая $3.83 + \log 2 \approx 4.52$, третья ≈ 4.92 и т.д, что и наблюдается на графиках.

4.4 Заключение главы

В этой главе была применена идея метода OBD к задаче тематического моделирования. Было оценено изменение логарифма правдоподобия (1.1) при занулении φ_{wt} или θ_{td} (Теорема 14). Для ускорения вычисления предложенных оценок и использования в алгоритме ARTM были предложены аппроксимации, которые вычисляются за приемлемое для практического применения время (Теорема 15).

Аппроксимации были реализованы в алгоритме ARTM и был проведён эксперимент на текстовой коллекции «20Newsgroups», сравнивающий предложенный метод с разреживанием с помощью регуляризатор разреживания ARTM (раздел 4.2). Эксперимент показал, что OBD позволяет добиться примерно того же уровня разреженности, но при этом, в отличие от подхода с регуляризатором, не увеличивает перплексию модели.

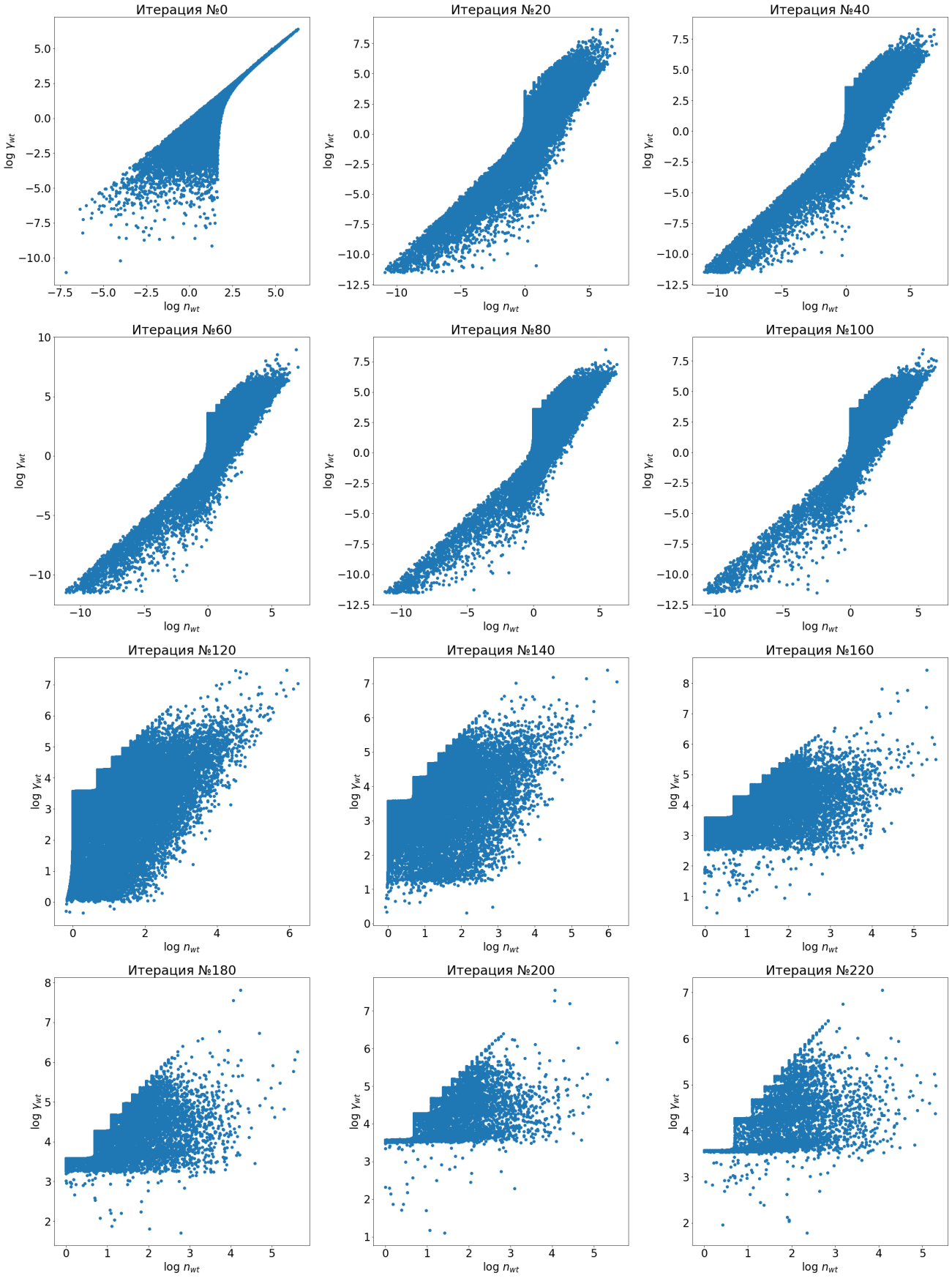

Рисунок 4.2 — Изменения структуры разреженности матрицы Φ на итерациях, $|T| = 10$.

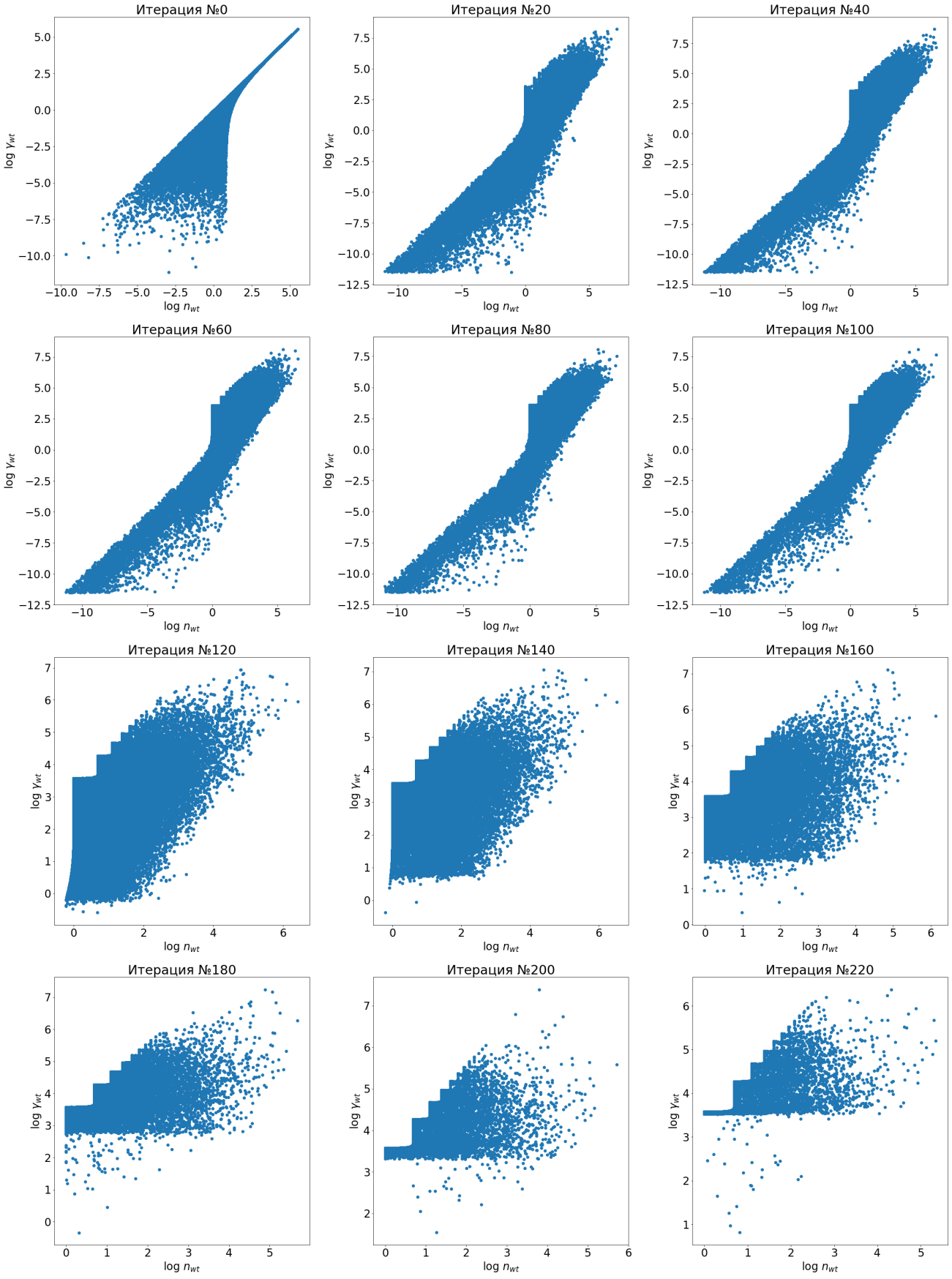

Рисунок 4.3 — Изменения структуры разреженности матрицы Φ на итерациях, $|T| = 25$.

Глава 5. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста

Важной особенностью тематических моделей является интерпретируемость: часто человек-эксперт может дать понятное описание каждой из компонент распределения φ_{wt}. Когерентность темы (topic coherence) — автоматическая мера качества тематических моделей, хорошо коррелирующая с асессорскими оценками интерпретируемости [55] (недостатком этого метода является то, что он основан на анализе ограниченного числа слов, составляющие маленькую долю от всей коллекции документов [56])

Генерация псевдодокументов — эффективный способ улучшения когерентности тем. Распространены подходы, основанные на со-встречаемостях: biterm topic model BTM [57]) и Word Network Topic Model WNTM [58]. Другой подход к построению псевдодокументов, особенно полезный для анализа коротких текстов, "склеивает" близкие документы вместе. Дальнейшее развитие подход псевдодокументов получил в работе [59], в которой комбинируются оптимизация в пространстве "слово-документ" и оптимизация в пространстве "слово-контекст".

Двумя существенными недостатками псевдодокументов являются необходимость настройки параметров (например, критерий склейки или размер контекстного окна) и необходимость большой предобработки данных (за счёт чего данный подход работает существенно дольше традиционных методов).

Другой способ улучшения качества, широко используемый в области машинного обучения в целом, — уменьшение числа параметров. Особенно релевантными в контексте тематического моделирования кажутся работы из области неотрицательного матричного разложения, показывающие полезность симметричности в задачах кластеризации [60—62].

При этом данные вопросы не освещены в литературе по тематическому моделированию. Насколько нам известно, используются лишь две техники для уменьшения числа параметров. Первая сводится к уменьшению числа скрытых тем (например, используя Иерархические Процессы Дирихле [63—66]). Вторая заключается в использовании разреженных распределений [26; 67].

В этой главе рассматривается применение симметричности к тематическим моделям (переход от разложения вида $M \approx X \cdot X^T$ к разложению вида

$M \approx X \cdot f(X)$) и предпринимается попытка обойтись без псевдодокументов за счёт более тщательного учёта информации, содержащейся в обычных документах.

5.1 Роль матрицы тем в документах и EM-алгоритм

С точки зрения вероятностного вывода матрица слова-темы (Φ) и матрица документы-темы (Θ) имеют одинаковую важность, поскольку обе являются скрытыми параметрами вероятностной модели (раздел 2.1.3). Однако, на практике, исследователи часто считают матрицу Φ более важной и относятся к матрице Θ как к чему-то вспомогательному, что может быть легко восстановлено из известных данных.

В первую очередь нужно упомянуть интерпретируемость. Интерпретируемость является желательным свойством хорошей тематической модели. Оценка интерпретируемости человеком обычно состоит из выбора небольшого набора самых вероятных слов для каждой темы и представления этого набора эксперту-человеку [55]. В этом процессе используется только матрица Φ.

Вторым примером подобного подхода является основополагающая работа [68], в которой измерялась интерпретируемость нескольких тематических моделей. Построенные модели вместе с использованной разметкой, были выложены в открытый доступ. Однако, была опубликована только матрица Φ (возможно, потому что авторы посчитали ее более ценной, или из-за неявного расчёта на то, что недостающее распределение Θ может быть восстановлено по выложенным данным).

К второстепенности матрицы Θ можно прийти и из практических соображений.

Во-первых, на практике часто встречаются задачи, требующие динамического расширения коллекции документов (например, анализ новостных потоков). Такое расширение может существенно увеличить $\mid D \mid$, практически не изменив размер словаря $\mid W \mid$ (что объясняется законом Хипса [69]).

Во-вторых, требование вычислительной эффективности естественным образом приводит к использованию параллельных, распределённых или онлайновых реализаций алгоритмов тематического моделирования. Наиболее

эффективным является следующий подход, реализованный в открытой библиотеке BigARTM: алгоритм разбивает входные данные на пакеты, которые обрабатываются разными потоками [70]. В результате, алгоритмы библиотеки никогда не хранят всю матрицу Θ, вместо этого элементы матрицы рассчитываются, когда они необходимы.

Для многих исследователей качество тематической модели эквивалентно прежде всего качеству матрицы Φ. Но с точки зрения самой тематической модели, Φ и Θ являются равноправными, и появление слов в документах коллекции объясняется при помощи обеих этих матриц. Качество матрицы Θ при этом никак не контролируется, поэтому тематическая модель может скомпенсировать "плохую" Φ специально подогнанной матрицей Θ.

Реализации тематического моделирования (особенно восстанавливающие элементы Θ "на лету") часто используют следующую эвристику: для получения θ_{td} конкретного документа d повторяются несколько итераций EM-алгоритма с фиксированной Φ. В этой процедуре вектор θ_{*d} сначала инициализируется некоторым образом (как правило, используется равномерное распределение), а затем итеративно обновляется по формуле $\theta_{td} \propto \sum_w n_{dw} p_{tdw}$ с пересчётом p_{tdw}. Обновление может происходить какое-то установленное количество итераций либо продолжаться до сходимости.

Отметим, что этот процесс может привести к переобучению, поскольку Θ целенаправленно оптимизируется для того, чтобы соответствовать заданной Φ. Кроме того, время обучения модели линейно зависит от числа этих итераций, и поэтому слишком большое их количество может существенно замедлить обучение.

Чтобы сделать роль Φ в EM-алгоритме более значимой, мы предлагаем заменить исходную оптимизационную задачу 1.4 на следующую:

$$L(\Phi, f(\Phi)) + R(\Phi, f(\Phi)) \to \max_{\Phi}, \tag{5.1}$$

где f — это некоторая функция, которая отображает матрицу темы-слова в матрицу документы-темы. Решение задачи 5.1 может отличаться от решения задачи 1.4.

5.2 Итерационный алгоритм для подхода ARTM без матрицы документы-темы

В этом разделе будет предложен итерационный алгоритм для оптимизационной задачи (5.1). Для этого на основе практических соображений будет выбрана функция f зависимости матриц Φ и Θ, и проведён вывод EM-алгоритма, аналогичный Теореме 1. Также результаты теорем о сходимости главы 2 будут перенесены на предложенный алгоритм.

5.2.1 Функция зависимости матриц документы-темы и темы-слова

Для дальнейшего изложения нужно определить функцию зависимости Φ и Θ. Другими словами, указать, как рассчитать вероятности тем в документах $p(t \mid d)$, зная только вероятности слов в темах $p(w \mid t)$.

Подчеркнём, что есть бесконечное множество возможных функций зависимости. Например, можно рассмотреть следующее бесконечное семейство: берётся какое-то начальное приближение для Θ, которое затем уточняется на протяжении $k, k \in \mathbb{N}$ итераций.

Однако, мы требуем, чтобы искомая функция была интерпретируемой, простой для анализа и лёгкой для вычислений. Естественным вариантом является усреднение распределений тем слов по всем словам, встречающимся в документе. Более формально:

$$P(t \mid d) \propto \sum_w n_{dw} P(t \mid w),$$

где $P(t \mid w)$ получены по формуле Байеса, предполагая, что распределение $p(t)$ равномерно:

$$P(t \mid w) = \frac{P(w \mid t)}{\sum_{s=1}^{T} P(w \mid s)} = \frac{\Phi_{wt}}{\sum_s \Phi_{ws}}.$$

Если мы обозначим $n_{dw}(\sum_w n_{dw})^{-1}$ за B_{dw}, то

$$\Theta_{td} = \sum_w B_{dw} \frac{\Phi_{wt}}{\sum_s \Phi_{ws}} \tag{5.2}$$

Эту формулу также можно проинтерпретировать как результат первой итерации процесса, описанного в 5.1.

5.2.2 Вывод EM-алгоритма

Для вывода EM алгоритма будут использоваться следующие обозначения:

$$A_{dw} = \frac{n_{dw}}{\sum_s \Phi_{ws}\Theta_{sd}}[n_{dw} > 0],$$
$$B_{dw} = \frac{n_{dw}}{\sum_w n_{dw}},$$
$$C_{dt} = (A\Phi)_{dt} + \frac{\partial R}{\partial \Theta_{td}},$$
$$h_w = \frac{1}{\sum_s \Phi_{ws}}.$$

Теорема 16. *В EM-алгоритме для 5.1 с формулой зависимости матрицы* Φ *и* Θ (5.2) *E-шаг останется без изменений, а M-шаг будет выглядеть следующим образом:*

$$\Phi_{wt}^{new} \propto \left(\sum_d n_{dw} p_{tdw} + \Phi_{wt}^{old} \left(\frac{\partial R}{\partial \Phi_{wt}} + h_w (C^T B)_{tw} - h_w^2 (\Phi^{old} C^T B)_{ww} \right) \right)_+ . \quad (5.3)$$

Доказательство.
Для вывода формул воспользуемся выводом обобщённого EM алгоритма (GEM). В GEM алгоритме на каждой итерации на E-шаге Φ и Θ фиксируются, считаются p_{tdw} и строится функционал (2.6):

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

Как было показано в доказательстве Теоремы 7, изменения функционала данного Q (значений между разными итерациями) являются нижней оценкой на изменения исходного регуляризированного логарифма правдоподобия. То есть, если $\Delta Q > 0$, то $\Delta(L + R) > 0$. Поэтому цель M-шага увеличить значение данного функционала по сравнению с Φ и Θ с предыдущей итерации.

То, что в предложенном новом подходе Θ — это функция от Φ, не меняет тот факт, что изменения Q — это нижняя оценка изменения $L + R$. Так как это

основное требование к функционалу на E-шаге, то поскольку оно выполняется, E-шаг нового алгоритма останется без изменений. Теперь цель M-шага — подобрать Φ, чтобы увеличить значение по сравнению с Φ с предыдущей итерации следующий функционал:

$$Q(\Phi, \Phi') = \sum_{dtw} n_{dw} p'_{tdw} \left(\ln \Phi_{wt} + \ln(\Theta(\Phi))_{td}\right) + R(\Phi, \Theta(\Phi)). \tag{5.4}$$

Найдём его производные:

$$\frac{\partial Q}{\partial \Phi_{vr}} = \frac{1}{\Phi_{vr}} \left(\sum_d n_{dv} p'_{rdv} + \Phi_{vr} \frac{\partial R}{\partial \Phi_{vr}} + \sum_{dtw} n_{dw} p'_{tdw} \frac{1}{\Theta_{td}} \frac{\partial \Theta_{td}}{\partial \Phi_{vr}} + \sum_{dt} \frac{\partial R}{\partial \Theta_{td}} \frac{\partial \Theta_{td}}{\partial \Phi_{vr}} \right).$$

Подставив вместо p'_{tdw} его выражение через Φ и Θ, получим:

$$\sum_w n_{dw} p_{tdw} \frac{1}{\Theta_{td}} = \sum_{dtw} n_{dw} \frac{\Phi_{wt}}{\sum_s \Phi_{ws} \Theta_{sd}} = \sum_{dtw} A_{dw} \Phi_{wt} = (A\Phi)_{dt},$$

то есть

$$C_{dt} = \frac{\sum_w n_{dw} p'_{tdw}}{\Theta_{td}} + \frac{\partial R}{\partial \Theta_{td}},$$

поэтому

$$\frac{\partial Q}{\partial \Phi_{vr}} = \frac{1}{\Phi_{vr}} \left(\sum_d n_{dv} p'_{rdv} + \Phi_{vr} \left(\frac{\partial R}{\partial \Phi_{vr}} + \sum_{dt} C_{dt} \frac{\partial \Theta_{td}}{\partial \Phi_{vr}} \right) \right).$$

Остаётся только найти $\dfrac{\partial \Theta_{td}}{\partial \Phi_{vr}}$.

$$\Theta_{td} = \sum_w B_{dw} \frac{\Phi_{wt}}{\sum_s \Phi_{ws}} = \sum_w B_{dw} \Phi_{wt} h_w.$$

$$\begin{aligned} \frac{\partial \Theta_{td}}{\partial \Phi_{vr}} &= \sum_w B_{dw}\ h_w \delta_{vwrt} + \sum_w B_{dw} \Phi_{wt} \frac{\partial h_w}{\partial \Phi_{vr}} = \\ &= \sum_w B_{dw} h_w \delta_{vwrt} - \sum_w B_{dw}\ \Phi_{wt}\ h_w^2\ \delta_{vw} = B_{dv}\ h_v\ \delta_{rt} - B_{dv}\ \Phi_{vt}\ h_w^2, \end{aligned}$$

где δ это символ Кронекера. Теперь

$$\begin{aligned} \sum_{dt} C_{dt} \frac{\partial \Theta_{td}}{\partial \Phi_{vr}} &= \sum_{dt} C_{dt} \left(B_{dv}\ h_v\ \delta_{rt} - B_{dv}\ \Phi_{vt}\ h_v^2 \right) = \\ &= h_v \sum_d C_{dr} B_{dv} - h_v^2 \sum_{dt} C_{dt} B_{dv} \Phi_{vt} = h_v (C^T B)_{rv} - h_v^2 (\Phi C^T B)_{vv}. \end{aligned}$$

Итого

$$\frac{\partial Q}{\partial \Phi_{vr}} = \frac{1}{\Phi_{vr}} \left(\sum_d n_{dv} p_{rdv} + \Phi_{vr} \left(\frac{\partial R}{\partial \Phi_{vr}} + h_v (C^T B)_{rv} - h_v^2 (\Phi C^T B)_{vv} \right) \right).$$

Далее, выписав условия Каруша-Куна-Таккера, по аналогии с Теоремой 1 получаем, что

$$\Phi_{vr}^{new} \propto \left(\sum_d n_{dv} p_{rdv} + \Phi_{vr}^{old} \left(\frac{\partial R}{\partial \Phi_{vr}} + h_v (C^T B)_{rv} - h_v^2 (\Phi^{old} C^T B)_{vv} \right) \right)_+ .$$

□

5.2.3 Анализ асимптотической сложности работы и сходимости алгоритма

На M-шаге предложенного алгоритма потребуется найти матрицы C, $C^T B$ и диагональ матрицы $\Phi C^T B$.

В силу разреженности A и B, умножение на эти матрицы может быть эффективно реализовано за $O(NS)$, где N – суммарная длина коллекции, а S — общая размерность умножаемых матриц. В данном случае это матрицы Φ и C, поэтому S равно числу тем.

Поскольку матрицы Φ и $C^T B$ уже подсчитаны, то диагональные элементы матрицы $\Phi C^T B$ могут быть найдены за $O(WT)$.

Так как вычисление p_{tdw} выполняется за такое же асимптотическое время, то изменение M-шага не приведёт с изменению асимптотики времени работы алгоритма.

Изменения функционала (5.4) являются нижней оценкой изменения $L+R$, поэтому будет верна следующая теорема:

Теорема 17. *Пусть регуляризатор R является дифференцируемой функцией при $\varphi_{wt}, \theta_{td} \in (0, 1]$, сохраняющей нуль, корректной, ε-разреживающей и δ-регулярной. Также допустим, что $Q(\Phi^{k+1}, \Phi^k) \geqslant Q(\Phi^k, \Phi^k)$ начиная с некоторой итерации k. Тогда последовательность p_{tdw}^k сходится в смысле дивергенции Кульбака–Лейблера для любых d и w таких, что $n_{dw} > 0$:*

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

Доказательство.
В доказательстве Теоремы 7 было показано

$$\Delta^k(L+R) = Q(\Phi^{k+1}, \Theta^{k+1}, \Phi^k, \Theta^k) - Q(\Phi^k, \Theta^k, \Phi^k, \Theta^k) + \mathrm{KL}\big(p_{dw}^k \,\big\|\, p_{dw}^{k+1}\big).$$

В предложенном алгоритме выполняется $\Theta^k = f(\Phi^k)$, поэтому это равенство всё ещё выполняется в терминах функционала (5.4):

$$\Delta^k(L+R) = Q(\Phi^{k+1}, \Phi^k) - Q(\Phi^k, \Phi^k) + \mathrm{KL}\big(p_{dw}^k \,\big\|\, p_{dw}^{k+1}\big).$$

При увеличении Q на итерациях будет выполнено

$$\Delta^k(L+R) \geqslant \mathrm{KL}\big(p_{dw}^k \,\big\|\, p_{dw}^{k+1}\big) \geqslant 0,$$

что будет влечь за собой сходимость параметров аналогично Теореме 7. □

Следствие 8. *Если в дополнение к условиям Теоремы 17 регуляризатор R сильно регулярен, а регуляризационная поправка рассчитывается не точке Φ^{old}, а в точке $\frac{n_{wt}}{n_t}$ по аналогии с (2.9), то*

$$|\varphi_{wt}^k - \varphi_{wt}^{k+1}| \to 0.$$

Доказательство.
Нетрудно заметить, что в предложенных условиях выполняются условия Следствия 1. □

5.3 Описание экспериментов с алгоритмом ARTM с быстрой векторизацией текста

Для оценки улучшения метрик качества тематических моделей от предложенного алгоритма была проведена серия экспериментов.

Эксперименты проводились на трёх стандартных текстовых коллекциях: 20Newsgroups (auto, motorcycles, baseball, hockey, crypt, electronics, med, space), NIPS Conference Papers 1987-2015 Data Set и Twitter Sentiment140 Data Set При построении моделей использовались $|T| = 25$ для 20Newsgroups, $|T| = 50$ для NIPS and Twitter.

Сравнивались несколько подходов в тематическом моделировании. `PLSA`, `LDA` и `sparse LDA` обозначают стандартные PLSA и LDA с сглаживающим и разреживающим значением параметра априорного распределения Дирихле. Два варианта предлагаемого подхода были рассмотрены: выведенный в Теореме 16 (`TARTM`) и популярный эвристический вывод `naive TARTM` где на каждой итерации Θ рассчитывается согласно (1.6) (вместо (5.3)) и Θ отбрасывается после расчёта p_{tdw}. Для проведения численных экспериментов на языке Python был реализован отдельный модуль, позволяющий проверять различные варианты реализации EM и ARTM.

Использование в качестве регуляризатора. Если сравнить формулы (1.6) и (5.3), то можно заметить, что они отличаются на слагаемое, имеющее такой же вид, как и $\frac{\partial R}{\partial \Phi_{wt}}$. Это означает, что вышеописанный итерационный процесс можно "сэмулировать" внутри традиционного подхода ARTM, введя фиктивный регуляризатор специального вида[1] и положив, что Θ должна получаться из Φ за одну итерацию EM-алгоритма (то есть по формуле (5.2)) [2].

Мы реализовали эту идею на практике, построив специальный регуляризатор внутри библиотеки TopicNet. TopicNet – открытая надстройка над библиотекой BigARTM, предоставляющая более удобные возможности по работе с пользовательскими регуляризаторами [32]. Наличие такого регуляризатора будет дополнительным фактором достоверности результатов эксперимента за счёт реализации в сторонней библиотеке и проверке на встроенной и поставляемой вместе с библиотекой текстовой коллекции 20NG.

Также это позволяет напрямую сравнить результаты TARTM и ARTM с традиционным набором регуляризаторов (сглаживание фоновых тем, разреживание предметных тем, декорреляция) и проверить взаимодействие предложенной формулы с дополнительными регуляризаторами. Для этой реализации сравнивались не значения метрик на итерациях, а финальные значения метрик после заданного числа итераций; мы положили параметр $\mid T \mid = 20$.

Использование для онлайн алгоритма. Предложенную модификацию можно использовать в онлайн алгоритме ARTM [30]. Поскольку в онлайн алгоритмах [30; 71; 72] данные поступают батчами из документов, то для них не хранится матрица Θ, а вычисляется «на лету» по матрице Φ и матрице

[1] Заметим, что это новое слагаемое внутри M-шага может не быть производной какой-либо функции.

[2] Библиотека BigARTM позволяет добиться последнего, если установить `num_document_passes = 1`.

данным батча. В эксперименте будут сравнены исходная версия онлайн алгоритма с вычислением Θ за одну итерацию по формуле (5.2) и онлайн версия алгоритма с быстрой векторизацией, в которой обновление (5.3) происходит по данным батча.

Для оценки качества полученных тематических моделей использовались следующие метрики:

Разреженность матрицы Φ. Разреженность определяется как доля нулей в матрице. Более высокая разреженность означает более высокую интерпретируемость и более низкое переобучение модели.

Средняя мера Жаккара между топ-словами. Это метрика показывает степень различности тем: средний коэффициент Жаккара равный 1 означает, что все найденные темы являются дубликатами друг друга, в то время как нулевой средний коэффициент говорит, что все темы уникальны. В эксперименте эта метрика рассчитывалась следующим способом:

$$\frac{1}{T(T-1)}\sum_{s\neq t}\frac{|K(t)\cap K(s)|}{|K(t)\cup K(s)|},$$

где $K(t)$ это множество, состоящее из 100 наиболее вероятных слов в теме t.

Среднее расстояние до ближайшей темы. Это другой способ оценки различности тем, анализирующий не только топ-слова темы, но всё распределение целиком. Для оценки этой метрики вычисляются попарные расстояния между всеми темами, для каждой темы определяется ближайшая (помимо неё самой) и эти расстояния усредняются. Здесь приведены расстояния, полученные на основе косинусной близости векторов (были рассмотрены и другие метрики близости, но качественной разницы в результатах не наблюдалось).

PPMI топ-слов. PMI между словами w и v определяется как

$$PMI(w,v)=\log\frac{|\,D\,|\,N(w,v)}{N(w)N(v)},$$

где $|D|$ — количество документов в коллекции, $N(w,v)$ — количество документов, где слова w и v содержатся одновременно, а $N(w)$ — количество документов, содержащих слов w. Чтобы избежать околонулевых значений, используется $PPMI=\max(PMI,0)$. Для расчёта метрики бралось множество из 30 наиболее вероятных слов темы и суммировался попарный PPMI по всем парам в этом множестве, а затем результат усреднялся по всем темам. Эта

метрика является широко используемым показателем согласованности тем и коррелирует с восприятием интерпретируемости тем человеком [55; 73; 74].

LogLift. Величина Lift была введена в работе [75], где рекомендовалось использовать её при показе темы пользователю (а именно, сортировать слова темы по Lift, а не по их вероятностям). $\mathrm{Lift}(w,t)$ определяется как отношение φ_{wt} к средней частоте слова w по коллекции. В недавней работе [76] была предложена метрика качества Log-Lift, основанная на этой величине: для 30 слов w_i, наиболее вероятных в теме t, вычисляется $\frac{1}{30}\sum_{i=1}^{30}\log \mathrm{Lift}(w_i,t)$, в дальнейшем усредняемая по всем темам. Было показано, что рассчитанная таким образом величина связана с долей неинформативных слов в темах, а также имеет существенную корреляцию с экспертными оценками качества тем.

Качество классификации. Чтобы оценить качество вычисления распределения тем в документах, для 20Newsgroups также оценивалось качество классификации документов по следующей схеме. Документы разделялись на три группы: train1, train2 и test. На train1 вместе с train2 обучалась тематическая модель и мы получали матрицу Θ для документов. Далее на этих признаках обучался SVM для многоклассовой классификации истинной метки документа. По кроссвалидации на train1 подбирались оптимальные параметры SVM (лучшее значение — cv fold). Затем оценивалось качество классификации на train2 (cv test). После чего оценивалось качество классификатора на преобразованных признаках документов из test. Чтобы получить матрицу Θ для новых документов требует сделать несколько итераций EM алгоритма, поэтому качество классификации оценивалось после каждой итерации. Качество классификации измерялось по accuracy.

Перплексия. Перплексия это стандартная метрика для вероятностных тематических моделей, она определяется по логарифму правдоподобия по следующей формуле:

$$perplexity(\Phi,\Theta) = \exp -\frac{L(\Phi,\Theta)}{n},$$

где $n = \sum_{d,w} n_{dw}$ — суммарная длина коллекции.

5.4 Результаты экспериментов с алгоритмом ARTM с быстрой векторизацией текста

TARTM	LDA
game team player play season hockey hit league fan baseball **last** run watch throw pitcher ball stat year sport score	game year team player get **last good** baseball win play **go** season hit fan **think** time **make well say** league
car bike buy engine sell speed drive price mile road ride owner dealer drive model driver motorcycle tire detector brake	car bike **get** engine buy new **also** drive mile **make** speed look tire **well** dealer brake wheel **go good** road
period st series vs playoff pt shot king canada ranger lead cup toronto play wing pittsburgh buffalo blue chicago round	period gm vs pt st chicago power pp april shot play buffalo pittsburgh islander flame series lead **first** scorer cup

Таблица 5 — 20newsgroups, примеры наиболее вероятных слов в темах. Слова общей лексики выделены жирным. TARTM убирает подобные слова из тем в отличие от LDA.

Алгоритм	Разреженность	Средняя мера Жаккара	PPMI	LogLift	Среднее расстояние до ближайшей темы
sparse LDA	0.896	0.044	1.570	0.503	0.587
smooth LDA	0	0.043	1.509	0.479	0.632
PLSA	0.869	0.050	1.517	0.459	0.586
ARTM + *Reg*	0.898	0.027	1.710	0.590	0.661
TARTM	0.893	0.007	1.716	0.952	0.895
TARTM + *Reg*	**0.929**	**0.003**	**1.788**	**1.020**	**0.953**

Таблица 6 — Результаты эксперимента с алгоритмом быстрой векторизации в TopicNet.

На Рисунках 5.1, 5.2 и 5.3 изображены основные результаты экспериментов. Во-первых, они показывают, что самые разреженные модели были ожидаемо получены разреживающим LDA. Тем не менее, TARTM даёт модели сравнимой разреженности, отдельно этого не оптимизируя. Во-вторых, модели TARTM демонстрируют наилучший результат по мере Жаккара (это можно

method	cv folds	cv test	1 iter	2 iter	3 iter	4 iter	5 iter	6 iter
lda, $T = 10$	0.684	0.687	0.59	0.672	0.697	0.704	0.704	0.703
tartm, $T = 10$	0.701	0.697	0.689	0.683	0.68	0.678	0.677	0.677
lda, $T = 25$	0.729	0.74	0.685	0.748	0.762	0.765	0.765	0.764
tartm, $T = 25$	0.768	0.77	0.752	0.749	0.747	0.746	0.746	0.746

Таблица 7 — 20newsgroups, качество классификации тематик по матрице Θ.

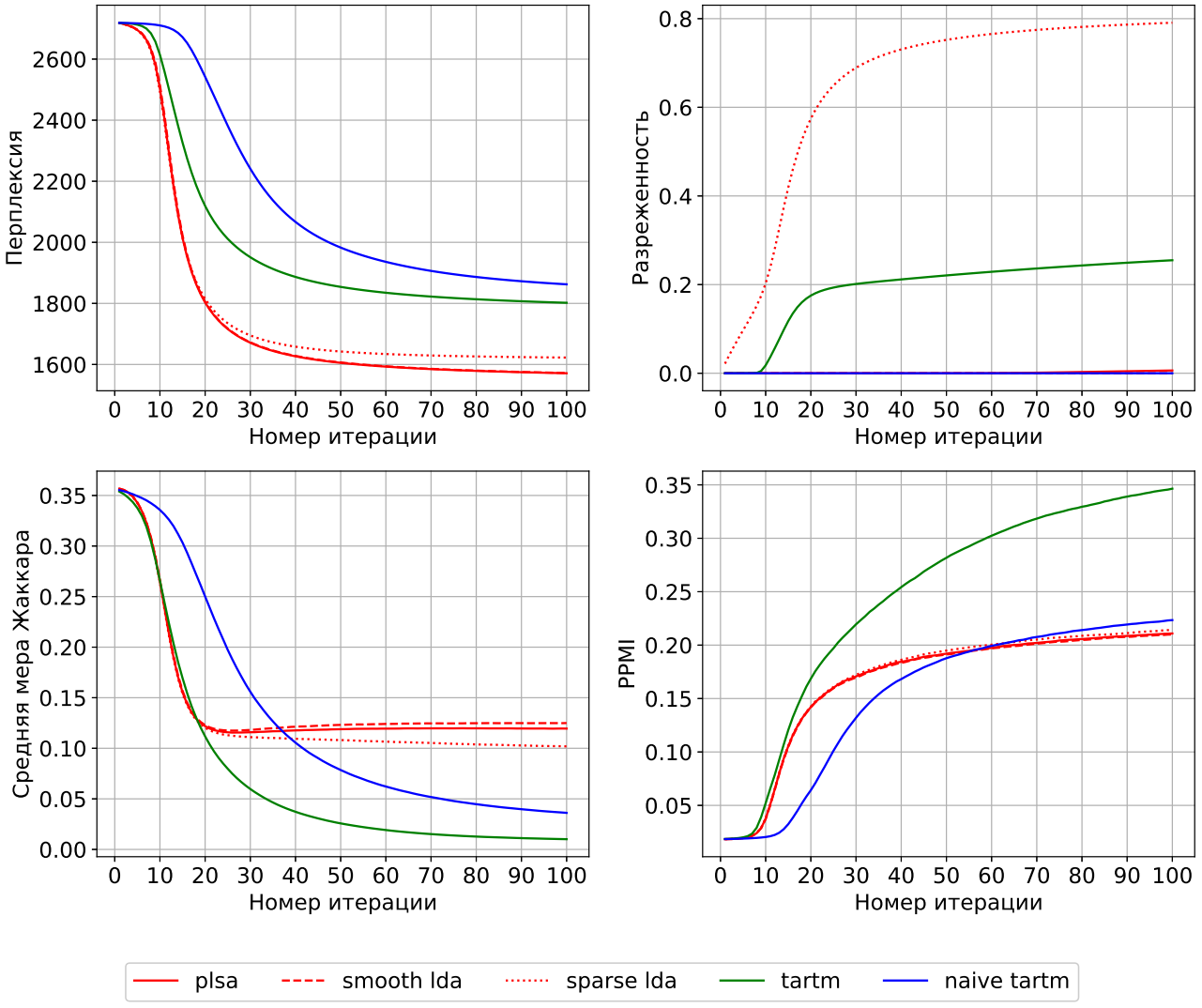

Рисунок 5.1 — Сравнение изменения метрик качества тематических моделей на итерациях для LDA и TARTM на коллекции NIPS, $|T| = 50$.

объяснить тем, что TARTM и LDA по-разному обрабатывают частые, но неинформативные слова; Таблица 5 демонстрирует это на примере трёх сравнимых тем).

В Таблице 6 приведены результаты, полученные через библиотеку TopicNet. Они подтверждают ранее описанные результаты, а также показыва-

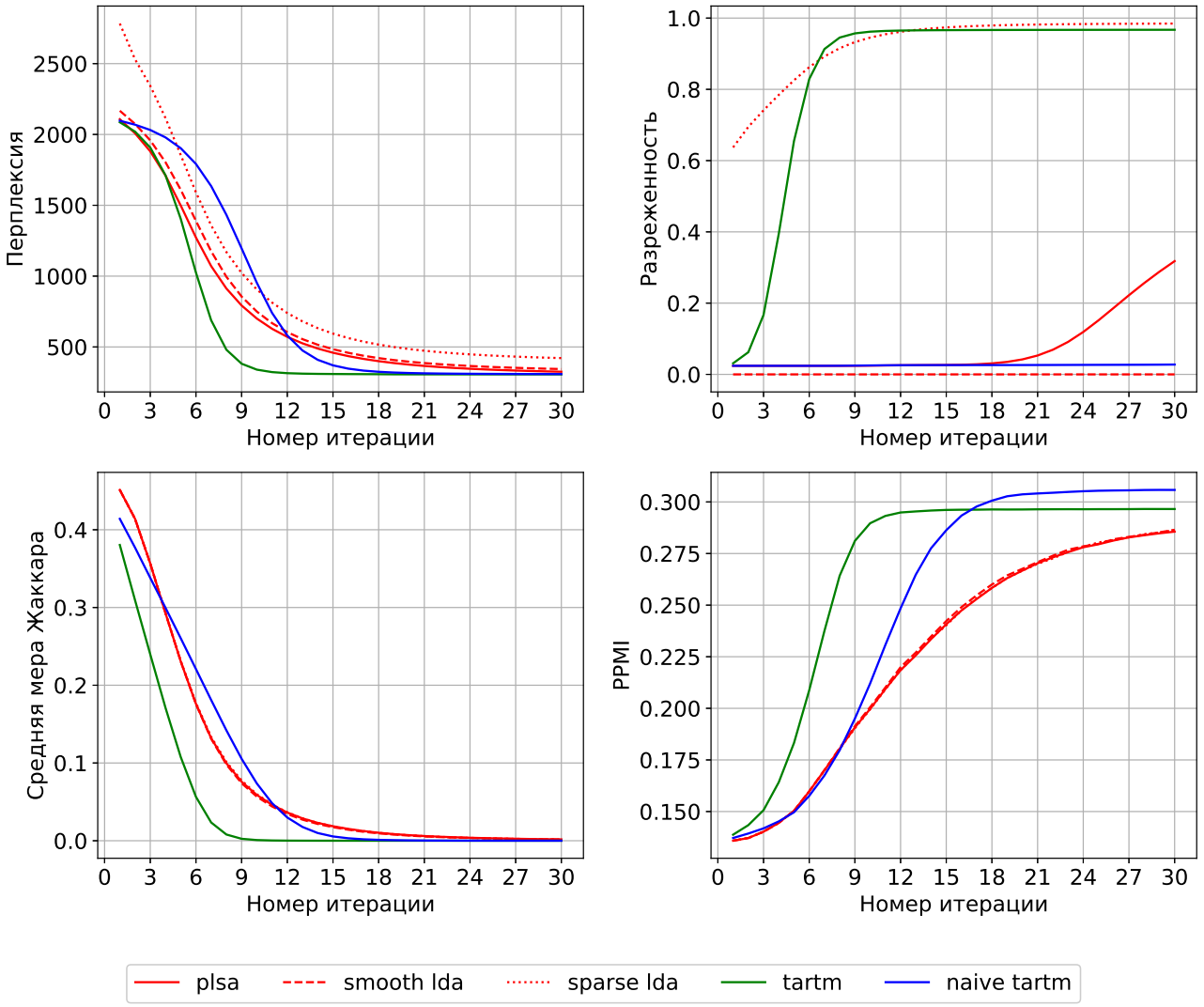

Рисунок 5.2 — Сравнение изменения метрик качества тематических моделей на итерациях для LDA и TARTM на коллекции Twitter, $|T| = 50$.

ют, что формула 5.3 эффективно комбинируется с другими регуляризаторами ARTM, что позволяет дополнительно увеличивать метрики качества.

Основное улучшение наблюдается в метриках PPMI и LogLift. TARTM превосходит все другие подходы на коллекциях NIPS и 20newsgroups, но уступает naive TARTM в наборе данных Twitter (однако TARTM сходится за меньшее количество итераций). Мы предполагаем, что это в основном из-за небольшого размера документа в коллекции Twitter, что делает оценку темы более восприимчивой к выбросам. В целом, мы видим, что отбрасывание матрицы Θ (наивным или строгим образом) дает более согласованные и интерпретируемые тематические модели.

В Таблице 7 приведены результаты по качеству классификации. Видно, что на кросс-валидации качество TARTM существенно выше, но на новых дан-

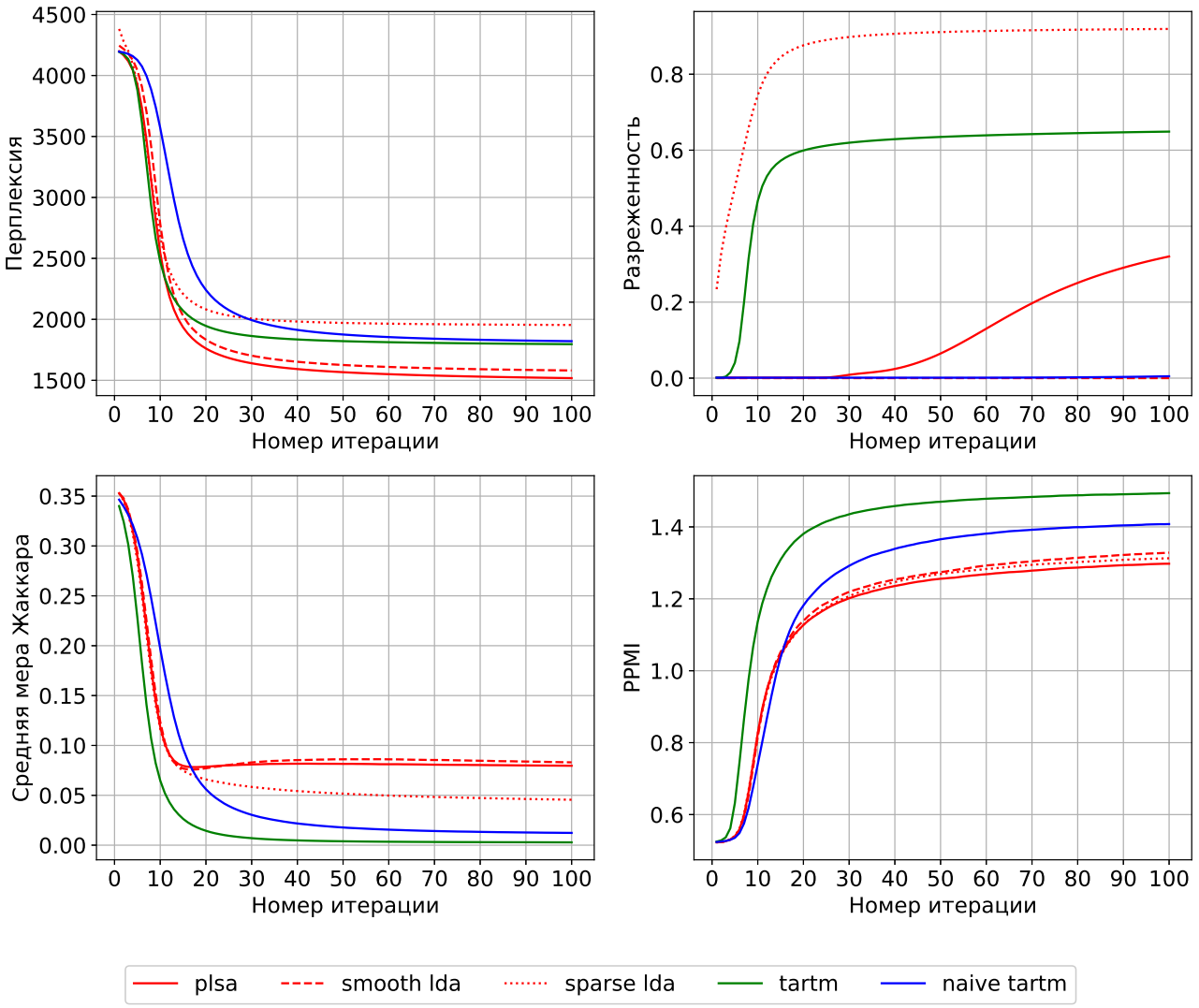

Рисунок 5.3 — Сравнение изменения метрик качества тематических моделей на итерациях для LDA и TARTM на коллекции 20Newsgroups, $|T| = 25$.

ных после нескольких итераций качество LDA становится выше, что означает, что LDA может подстроиться под новые данные. Тем не менее, на первых итерациях качество классификации у TARTM выше и оно сохраняется на всех итерациях, это свидетельствует в пользу того, что TARTM не подстраиваются под данные, а с первой же итерации показывают финальное качество классификации.

Основное объяснение полученных результатов следующее. PLSA и LDA предсказывают появление слов в документах как с помощью матрицы Φ, так и с помощью матрицы Θ, в то время как TARTM использует только матрицу Φ. Это означает, что PLSA и LDA могут “скорректироовать” недостатки матрицы Φ за счёт правильного подбора матрицы Θ, а TARTM может “исправлять” эти недостаки только меняя саму Φ.

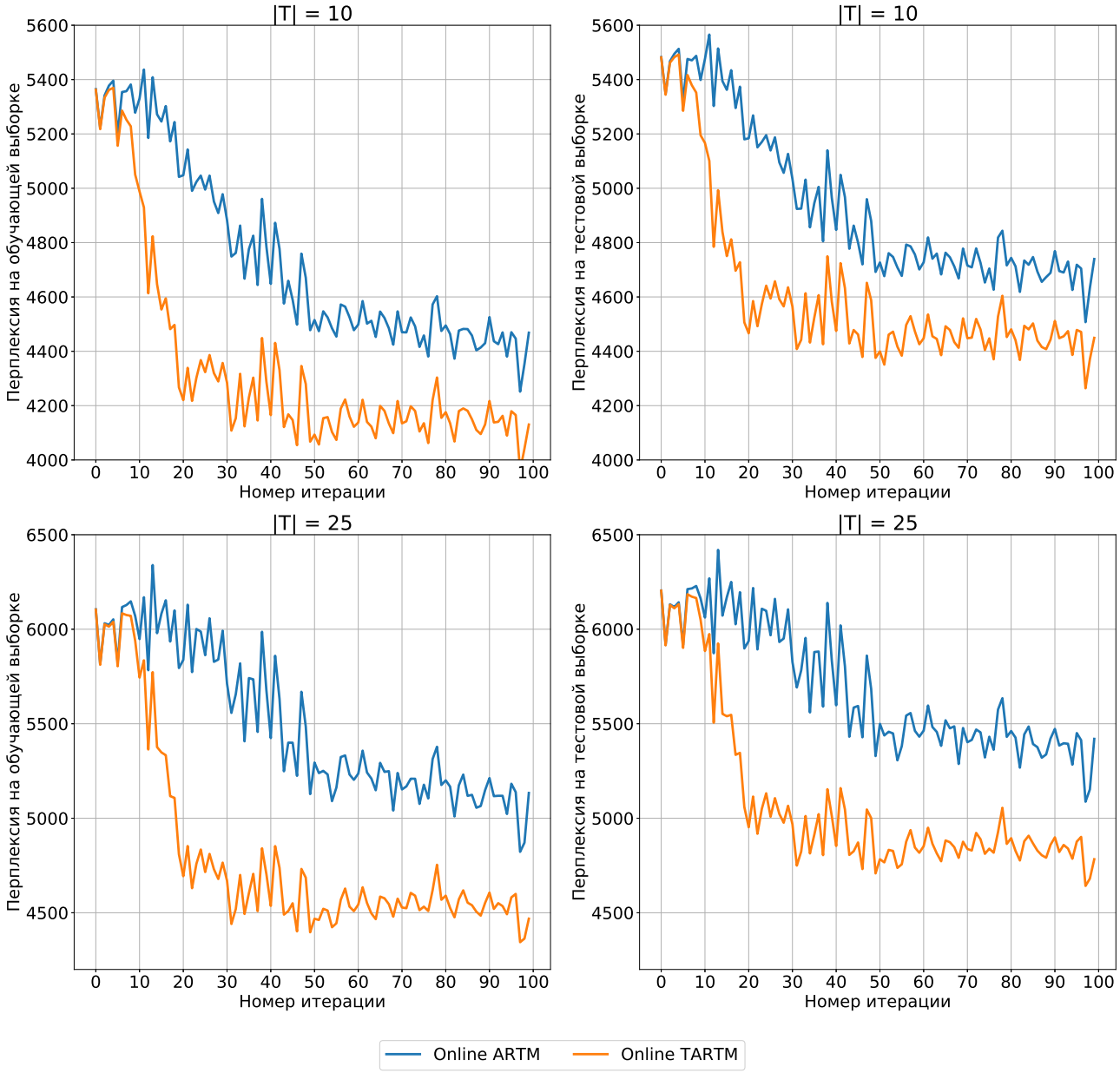

Рисунок 5.4 — Сравнение перплексии оригинального онлайн алгоритма ARTM и онлайн версии алгоритма TARTM.

Приведём простой пример, иллюстрирующий данные рассуждения. Допустим, коллекция состоит из 7 слов и 6 документов:

```
medicine and spices
herbs and spices
herbs and spices and chicken
honey and spices
medicine and herbs
medicine and honey
```

Как могла бы выглядеть “хорошая” тематическая модель из 4 тем, построенная на этой коллекции? Неинформативное слово "and" может либо лежать в какой-то единственной “фоновой” теме ($\exists t_0 : \varphi_{wt_0} > 0, \varphi_{w*} = 0$ для $w =$"and"), либо распределиться между несколькими “информативными” темами. Первый вариант кажется более естественным. Из 1000 запусков PLSA и TARTM с разными начальными приближениями PLSA ни разу не выделил "and" в отдельную тему, в то время как TARTM сделал это в 365 случаях. При этом матрица Θ, полученная в TARTM содержала от 3 до 7 нулей, а в PLSA от 12 до 17. Это показывает как PLSA с помощью нулей в матрице Θ “прячет” недостатки, вызванные шумами в виде наличия "and" во всех темах.

5.5 Заключение главы

В этой главе была предложена модификация оптимизационной задачи (1.4), которая уменьшает количество оптимизируемых параметров и повышает уникальность и когерентность получаемых тем. Предложенный в Теореме 16 алгоритм не увеличивает вычислительную сложность или количество необходимых обучающих примеров. Также результаты о сходимости алгоритма ARTM из главы 2 были перенесены на предложенный алгоритм (Теорема 17). Эксперименты на реальных данных показывают, что предложенный алгоритм действительно улучшает качество тем (раздел 5.4).

Важным аспектом предлагаемого алгоритма является его совместимость с подходом ARTM, что позволяет включать произвольное количество дополнительных регуляризаторов, чтобы точно настроить решение поставленной задачи.

Открытым вопросом остаётся адаптация предложенного подхода к мультимодальным тематическим моделям [30], в которых имеется несколько матриц Φ

Заключение

Основные результаты работы заключаются в следующем.

1. Теорема о достаточных условиях сходимости алгоритма ARTM.
2. Теорема о достаточных условиях единственности стохастического матричного разложения.
3. Модификация алгоритма ARTM, ускоряющая сходимость итерационного процесса.
4. Метод разреживания тематической модели, не увеличивающий перплексию получаемой модели.

Список сокращений и условных обозначений

NLP	Natural Language Processing
NLU	Natural Language Understanding
PTM	Probabilistic Topic Model
PLSA	Probabilistic Latent Semantic Analysis
LDA	Latent Dirichlet Allocation
ARTM	Additive Regularization of Topic Models
EM	Expectation Maximization
GEM	Generalized Expectation Maximization
tSNE	t-Distributed Stochastic Neighbor Embedding
OBD	Optimal Brain Damage
BTM	Biterm Topic Model
WNTM	Word Network Topic Model
TARTM	Thetaless ARTM

Список литературы

1. Distributed Representations of Words and Phrases and their Compositionality [Текст] / T. Mikolov [и др.] // Advances in Neural Information Processing Systems 26 / под ред. C. J. C. Burges [и др.]. — Curran Associates, Inc., 2013. — С. 3111—3119. — URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.
2. *Potapenko, A.* Interpretable Probabilistic Embeddings: Bridging the Gap Between Topic Models and Neural Networks [Текст] / A. Potapenko, A. Popov, K. Vorontsov // Communications in Computer and Information Science. — Springer International Publishing, 11.2017. — С. 167—180. — URL: https://doi.org/10.1007/978-3-319-71746-3%5C_15.
3. *Wang, W.* Instantaneous Versus Convolutive Non-Negative Matrix Factorization [Текст] / W. Wang. — 2011.
4. *Feng, Y.* Topic Models for Image Annotation and Text Illustration [Текст] / Y. Feng, Lapata, Mirella. — 2010.
5. *Hospedales, T.* Video Behaviour Mining Using a Dynamic Topic Model [Текст] / T. Hospedales, S. Gong, T. Xiang // International Journal of Computer Vision. — 2011. — Дек. — Т. 98, № 3. — С. 303—323.
6. Simultaneous image classification and annotation based on probabilistic model [Текст] / X.-x. LI [и др.] // The Journal of China Universities of Posts and Telecommunications. — 2012. — Апр. — Т. 19, № 2. — С. 107—115.
7. *Pritchard J. K. Stephens M., D. P.* Inference of population structure using multilocus genotype data [Текст] / D. P. Pritchard J. K. Stephens M. // Genetics. — 2000. — Т. 155. — С. 945—959.
8. Multi-view methods for protein structure comparison using latent dirichlet allocation [Текст] / S. Shivashankar [и др.] // Bioinformatics. — 2011. — Июнь. — Т. 27, № 13. — С. i61—i68.
9. *Vulić, I.* Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora [Текст] / I. Vulić, W. D. Smet, M.-F. Moens // Information Retrieval. — 2012. — Май. — Т. 16, № 3. — С. 331—368.

10. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications [Текст] / I. Vulić [и др.] // Information Processing & Management. — 2015. — Янв. — Т. 51, № 1. — С. 111—147.

11. *Ianina, A.* Multi-objective Topic Modeling for Exploratory Search in Tech News [Текст] / A. Ianina, L. Golitsyn, K. Vorontsov // Communications in Computer and Information Science. — Springer International Publishing, 11.2017. — С. 181—193. — URL: https://doi.org/10.1007/978-3-319-71746-3%5C_16.

12. *Ianina, A.* Regularized Multimodal Hierarchical Topic Model for Document-by-Document Exploratory Search [Текст] / A. Ianina, K. Vorontsov //. — 11.2019. — С. 131—138.

13. *Nikolenko, S. I.* SVD-LDA: Topic Modeling for Full-Text Recommender Systems [Текст] / S. I. Nikolenko // Proc. 14th Mexican International Conference on Artificial Intelligence. Т. 9414. — Springer, 2015. — С. 67—79. — (Lecture Notes in Computer Science).

14. *Nikolenko, S. I.* Topic Modelling for Qualitative Studies [Текст] / S. I. Nikolenko, S. Koltcov, O. Koltsova // Journal of Information Science. — Thousand Oaks, CA, USA, 2017. — Т. 43, № 1. — С. 88—102.

15. *Pan, C.* Research paper recommendation with topic analysis [Текст] / C. Pan, W. Li // 2010 International Conference On Computer Design and Applications. Т. 4. — IEEE. 2010. — С. 264—268.

16. *Hofmann, T.* Probabilistic latent semantic indexing [Текст] / T. Hofmann // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — ACM. 1999. — С. 50—57.

17. *Blei, D. M.* Latent dirichlet allocation [Текст] / D. M. Blei, A. Y. Ng, M. I. Jordan // the Journal of machine Learning research. — 2003. — Т. 3. — С. 993—1022.

18. Applications of topic models [Текст] / J. Boyd-Graber, Y. Hu, D. Mimno [и др.] // Foundations and Trends® in Information Retrieval. — 2017. — Т. 11, № 2/3. — С. 143—296.

19. *Cohn, D.* The missing link-a probabilistic model of document content and hypertext connectivity [Текст] / D. Cohn, T. Hofmann // Advances in neural information processing systems. — 2001. — С. 430—436.

20. *McCallum, A.* The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email [Текст] / A. McCallum, A. Corrada-Emmanuel, X. Wang. — 2005.

21. *Nallapati, R.* Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs [Текст] / R. Nallapati, W. W. Cohen // ICWSM. — 2008.

22. Probabilistic author-topic models for information discovery [Текст] / M. Steyvers [и др.] // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2004. — С. 306—315.

23. *Zosa, E.* Multilingual Dynamic Topic Model [Текст] / E. Zosa, M. Granroth-Wilding // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — Varna, Bulgaria : INCOMA Ltd., 09.2019. — С. 1388—1396. — URL: https://www.aclweb.org/anthology/R19-1159.

24. *Gruber, A.* Hidden topic Markov models [Текст] / A. Gruber, Y. Weiss, M. Rosen-Zvi // International conference on artificial intelligence and statistics. — 2007. — С. 163—170.

25. *Wallach, H. M.* Topic modeling: beyond bag-of-words [Текст] / H. M. Wallach // Proceedings of the 23rd international conference on Machine learning. — ACM. 2006. — С. 977—984.

26. *Vorontsov, K.* Additive regularization of topic models [Текст] / K. Vorontsov, A. Potapenko // Machine Learning. — 2015. — Т. 101, № 1—3. — С. 303—323.

27. *Vorontsov, K.* Additive regularization of topic models [Текст] / K. Vorontsov, A. Potapenko // Machine Learning. — 2014. — Дек. — Т. 101, № 1—3. — С. 303—323. — URL: https://doi.org/10.1007/s10994-014-5476-6.

28. *Vorontsov, K.* Additive Regularization of Topic Models for Topic Selection and Sparse Factorization [Текст] / K. Vorontsov, A. Potapenko, A. Plavin // Statistical Learning and Data Sciences. — Springer International Publishing, 2015. — С. 193—202. — URL: https://doi.org/10.1007/978-3-319-17091-6_14.

29. Fast and modular regularized topic modelling [Текст] / D. Kochedykov [и др.] // 2017 21st Conference of Open Innovations Association (FRUCT). — IEEE, 11.2017. — URL: https://doi.org/10.23919/fruct.2017.8250181.

30. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections [Текст] / K. Vorontsov [и др.] // Communications in Computer and Information Science. — Springer International Publishing, 2015. — С. 370—381. — URL: https://doi.org/10.1007/978-3-319-26123-2_36.

31. *Frei, O.* Parallel Non-blocking Deterministic Algorithm for Online Topic Modeling [Текст] / O. Frei, M. Apishev // Communications in Computer and Information Science. — Springer International Publishing, 2017. — С. 132—144. — URL: https://doi.org/10.1007/978-3-319-52920-2_13.

32. TopicNet: Making Additive Regularisation for Topic Modelling Accessible [Текст] / V. Bulatov [и др.] // Proceedings of The 12th Language Resources and Evaluation Conference. — 2020. — С. 6745—6752.

33. *Hadamard, J.* Sur les Problèmes aux Dérivées Partielles et Leur Signification Physique [Текст] / J. Hadamard // Princeton University Bulletin. Т. 13. — 1902. — С. 49—52.

34. *Tikhonov, A. N.* Solutions of ill-posed problems [Текст] / A. N. Tikhonov, V. I. Arsenin. — Winston ; distributed solely by Halsted Press Washington : New York, 1977. — xiii, 258 p. :

35. *Dempster, A. P.* Maximum likelihood from incomplete data via the EM algorithm [Текст] / A. P. Dempster, N. M. Laird, D. B. Rubin // Journal of the royal statistical society. Series B (methodological). — 1977. — С. 1—38.

36. *Wu, C. J.* On the convergence properties of the EM algorithm [Текст] / C. J. Wu // The Annals of statistics. — 1983. — С. 95—103.

37. *Donoho, D.* When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts? [Текст] / D. Donoho, V. Stodden. — 2004.

38. Theorems on Positive Data: On the Uniqueness of NMF [Текст] / H. Laurberg [и др.] // Computational Intelligence and Neuroscience. — 2008. — Т. 2008. — С. 1—9.

39. *Gillis, N.* Sparse and unique nonnegative matrix factorization through data preprocessing [Текст] / N. Gillis // The Journal of Machine Learning Research. — 2012. — Т. 13, № 1. — С. 3349—3386.

40. *Селезнев, Н.* Автоматическое извлечение атрибутов водителя из логов мобильного приложения такси [Текст] / Н. Селезнев, И. Ирхин, В. Кантор // Труды МФТИ. — 2018. — Т. 10, № 3. — С. 5—15.

41. *Ирхин, И.* Сходимость алгоритма аддитивной регуляризации тематических моделей [Текст] / И. Ирхин, К. Воронцов // Труды Института математики и механики УрО РАН. — 2020. — Т. 26, № 3. — С. 57—68.

42. *Дербаносов, Р. Ю.* Проблемы устойчивости и единственности стохастического матричного разложения [Текст] / Р. Ю. Дербаносов, И. Ирхин // Журнал вычислительной математики и математической физики. — 2020. — Т. 60, № 3. — С. 19—28.

43. *Ирхин, И.* Аддитивная регуляризация тематических моделей с быстрой векторизацией текста [Текст] / И. Ирхин, В. Булатов, К. Воронцов // Компьютерные исследования и моделирование. — 2020. — Т. 12, № 6.

44. *Vorontsov, K.* Additive regularization for topic models of text collections [Текст] / K. Vorontsov // Doklady Mathematics. Т. 89. — Citeseer. 2014. — С. 301—304.

45. *Vorontsov, K.* Tutorial on probabilistic topic modeling: additive regularization for stochastic matrix factorization [Текст] / K. Vorontsov, A. Potapenko // Analysis of Images, Social networks and Texts. — Springer, 2014. — С. 29—46.

46. *Zangwill, W. I.* Convergence conditions for nonlinear programming algorithms [Текст] / W. I. Zangwill // Management Science. — 1969. — Т. 16, № 1. — С. 1—13.

47. *Topsøe, F.* Some inequalities for information divergence and related measures of discrimination [Текст] / F. Topsøe // Information Theory, IEEE Transactions on. — 2000. — Т. 46, № 4. — С. 1602—1609.

48. *Lang, K.* 20 Newsgroups [Текст] / K. Lang. — 2008. — Data retrieved from the dataset's official website, http://qwone.com/~jason/20Newsgroups/.

49. *Tan, Y.* Topic-weak-correlated Latent Dirichlet allocation [Текст] / Y. Tan, Z. Ou // 2010 7th International Symposium on Chinese Spoken Language Processing. — IEEE, 11.2010. — URL: https://doi.org/10.1109/iscslp.2010.5684906.

50. *Apishev, M.* Learning Topic Models with Arbitrary Loss [Текст] / M. Apishev, K. Vorontsov // 2020 26th Conference of Open Innovations Association (FRUCT). — IEEE, 04.2020. — URL: https://doi.org/10.23919/fruct48808.2020.9087559.

51. *Vorontsov, K. V.* Additive regularization for topic models of text collections [Текст] / K. V. Vorontsov // Doklady Mathematics. — 2014. — Май. — Т. 89, № 3. — С. 301—304.

52. *Vorontsov, K.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization [Текст] / K. Vorontsov, A. Potapenko. — 2014.

53. *Maaten, L. van der.* Viualizing data using t-SNE [Текст] / L. van der Maaten, G. Hinton // Journal of Machine Learning Research. — 2008. — Нояб. — Т. 9. — С. 2579—2605.

54. *Cun, Y. L.* Optimal Brain Damage [Текст] / Y. L. Cun, J. S. Denker, S. A. Solla // Advances in Neural Information Processing Systems 2. — San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1990. — С. 598—605.

55. *Röder, M.* Exploring the space of topic coherence measures [Текст] / M. Röder, A. Both, A. Hinneburg // Proceedings of the eighth ACM international conference on Web search and data mining. — ACM. 2015. — С. 399—408.

56. *Alekseev, V.* Intra-text coherence as a measure of topic models' interpretability [Текст] / V. Alekseev, V. Bulatov, K. Vorontsov // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue. — 2018. — С. 1—13.

57. A biterm topic model for short texts [Текст] / X. Yan [и др.] // Proceedings of the 22nd international conference on World Wide Web - WWW '13. — ACM Press, 2013. — URL: https://doi.org/10.1145/2488388.2488514.

58. *Zuo, Y.* Word network topic model: a simple but general solution for short and imbalanced texts [Текст] / Y. Zuo, J. Zhao, K. Xu // Knowledge and Information Systems. — 2015. — Сент. — Т. 48, № 2. — С. 379—398. — URL: https://doi.org/10.1007/s10115-015-0882-z.

59. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations [Текст] / T. Shi [и др.] // Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18. — ACM Press, 2018. — URL: https://doi.org/10.1145/3178876.3186009.

60. Symmetric Nonnegative Matrix Factorization: Algorithms and Applications to Probabilistic Clustering [Текст] / Z. He [и др.] // IEEE Transactions on Neural Networks. — 2011. — Дек. — Т. 22, № 12. — С. 2117—2131. — URL: https://doi.org/10.1109/tnn.2011.2172457.

61. *Kuang, D.* Symmetric Nonnegative Matrix Factorization for Graph Clustering [Текст] / D. Kuang, C. Ding, H. Park // Proceedings of the 2012 SIAM International Conference on Data Mining. — Society for Industrial, Applied Mathematics, 04.2012. — URL: https://doi.org/10.1137/1.9781611972825.10.

62. *Mao, X.* On Mixed Memberships and Symmetric Nonnegative Matrix Factorizations [Текст] / X. Mao, P. Sarkar, D. Chakrabarti // Proceedings of the 34th International Conference on Machine Learning. Т. 70 / под ред. D. Precup, Y. W. Teh. — International Convention Centre, Sydney, Australia : PMLR, 06–11 Aug.2017. — С. 2324—2333. — (Proceedings of Machine Learning Research). — URL: http://proceedings.mlr.press/v70/mao17a.html.

63. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes [Текст] / Y. W. Teh [и др.] // Advances in Neural Information Processing Systems 17 / под ред. L. K. Saul, Y. Weiss, L. Bottou. — MIT Press, 2005. — С. 1385—1392. — URL: http://papers.nips.cc/paper/2698-sharing-clusters-among-related-groups-hierarchical-dirichlet-processes.pdf.

64. *Blei, D.* Probabilistic Topic Models [Текст] / D. Blei, L. Carin, D. Dunson // IEEE Signal Processing Magazine. — 2010. — Нояб. — URL: https://doi.org/10.1109/msp.2010.938079.

65. *Boyd-graber, J. L.* Syntactic Topic Models [Текст] / J. L. Boyd-graber, D. M. Blei // Advances in Neural Information Processing Systems 21 / под ред. D. Koller [и др.]. — Curran Associates, Inc., 2009. — С. 185—192. — URL: http://papers.nips.cc/paper/3398-syntactic-topic-models.pdf.

66. *Chirkova, N. A.* Additive Regularization for Hierarchical Multimodal Topic Modeling [Текст] / N. A. Chirkova, K. V. Vorontsov // Journal Machine Learning and Data Analysis. — 2016. — Т. 2, № 2. — С. 187—200.

67. The dual-sparse topic model: mining focused topics and focused terms in short text [Текст] / T. Lin [и др.] // Proceedings of the 23rd international conference on World wide web. — ACM. 2014. — С. 539—550.

68. Reading tea leaves: How humans interpret topic models [Текст] / J. Chang [и др.] // Advances in neural information processing systems. — 2009. — С. 288—296.

69. *Powers, D. M. W.* Applications and Explanations of Zipf's Law [Текст] / D. M. W. Powers // New Methods in Language Processing and Computational Natural Language Learning. — 1998. — URL: https://www.aclweb.org/anthology/W98-1218.

70. *Frei, O.* Parallel non-blocking deterministic algorithm for online topic modeling [Текст] / O. Frei, M. Apishev // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2016. — С. 132—144.

71. *Bassiou, N.* Online PLSA: Batch Updating Techniques Including Out-of-Vocabulary Words [Текст] / N. Bassiou, C. Kotropoulos // IEEE Transactions on Neural Networks and Learning Systems. — 2014. — Февр.

72. *Hoffman, M.* Online Learning for Latent Dirichlet Allocation [Текст] / M. Hoffman, D. Blei, F. Bach //. Т. 23. — 11.2010. — С. 856—864.

73. *Lau, J. H.* Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality [Текст] / J. H. Lau, D. Newman, T. Baldwin // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. — 2014. — С. 530—539.

74. Automatic evaluation of topic coherence [Текст] / D. Newman [и др.] // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Association for Computational Linguistics. 2010. — С. 100—108.

75. *Taddy, M.* On estimation and selection for topic models [Текст] / M. Taddy // Artificial Intelligence and Statistics. — 2012. — С. 1184—1193.

76. *Fan, A.* Assessing topic model relevance: Evaluation and informative priors [Текст] / A. Fan, F. Doshi-Velez, L. Miratrix // Statistical Analysis and Data Mining: The ASA Data Science Journal. — 2019. — Т. 12, № 3. — С. 210—222.

Список рисунков

Список таблиц