Министерство образования и науки Российской Федерации

Государственное образовательное учреждение
высшего профессионального образования
«МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)»

Факультет Инноваций и высоких технологий

Кафедра Анализа данных

Тип работы:
Магистерская диссертация по направлению
01.04.02 «Прикладная математика и информатика»
Номер программы 010400

Тема работы:

Сходимость численных методов вероятностного тематического моделирования

Научный руководитель:
_____ д.ф.-м.н., профессор К. В. Воронцов

Работу выполнил:
Студент 093 группы
_____ И. А. Ирхин

Москва 2016

Содержание

1 Введение

Тематическое моделирование — одно из современных приложений машинного обучения к анализу текстов, активно развивающееся с конца 90-х годов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.

Вероятностная тематическая модель описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем. Предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из смеси таких распределений, и ставится задача восстановления компонент смеси по выборке.

Одним из основных приложений тематического моделирования является информационный поиск [1]. Современные поисковые алгоритмы основаны на инвертированных индексах, за счёт которых система ищет документы, содержащие все слова запроса, поэтому скорее всего будет найдено очень мало документов. Данная проблема решается при помощи тематического (разведочного) поиска. Основная его идея состоит в построении тематической модели текста запроса, извлечении тем и дальнейшего поиска по данным темам. Фактически, для поиска документов используются те же механизмы обратного индексирования, только словами считаются темы, представленные в документах. Новые технологии информационного поиска на основе тематического моделирования в настоящее время активно развиваются [3, 5].

Тематическое моделирование также применяется для определения трендов в новостных потоках или научных работах [8, 23], для многоязычного информационного поиска [29, 31], при анализе структур социальных сетей [20, 25], в задачах классификации, кластеризации и категоризации документов [22, 34], для тематической сегментации текстов [32], при построении рекомендательных систем [16, 18].

1.1 О методах вероятностного тематического моделирования

Для задачи вероятностоного тематического моделирования классическим решением является Вероятностный Латентный Семантический Анализ (Probabilistic Latent Semantic Analysis, PLSA). Модель была предложена Томасом Хофманном в 1999 году [12]. Одним из недостатков этой модели является то, что она задаёт закон порождения слов в документах, но не закон порождения самих документов [15]. Также неясно, как оценивать слова, ранее не встречавшиеся в коллекции. Поэтому в 2003 году была предложена модель Латентного Размещения Дирихле (Latent Dirichlet Allocation, LDA) [4], лишённая данных недостатков. В LDA предполагается, что каж-

дое слово в документе порождено некоторой латентной темой, при этом в явном виде моделируется распределение слов в каждой теме, а также априорное распределение тем в документе. С тех пор было предложено множество моделей, базирующихся на данном подходе, которые постепенно усложняли вероятностную модель за счёт учёта связей между документами [6, 17, 19], метаданных о документах [21] и информации о порядке слов в документе [11, 30].

Данные модели объединяет общая схема. Сначала вводится вероятностная модель коллекции документов, а затем оценивается апостериорное распределение параметров. Для этого существует набор основных алгоритмов. Ниже приводится их краткий обзор.

EM-алгоритм (Expectation-Maximization) [2] обычно используется для точного оценивания параметров. Например, он используется для решения задачи PLSA. Основными недостатками данного метода являются быстрый рост количества параметров от числа слов, тем и документов, а также большое число локальных максимумов у функции правдоподобия. Более того, точное оценивание апостериорных распределений скрытых параметров может быть вычислительно неэффективным. В этом случае используются вариационные методы [14], которые являются специальной версией EM-алгоритма. Они позволяют оценивать узкие нижние и верхние доверительные границы для значений скрытых переменных в наблюдаемом документе. Эти методы обладают теми же недостатками, что и исходный EM-алгортим.

Методы Монте-Карло для марковских цепей (Markov chain Monte Carlo, MCMC) [9, 13] широко используются как эффективные приближенные процедуры генерации значений из распределений высоких размерностей. Одной из таких процедур является сэмплирование Гиббса [10]. Оно применяется, когда вычисление или хранение функции распределения слишком ресурсоёмко, а генерация случайной выборки из этого распределения — нет. Тогда исходное распределение заменяется несмещённой эмпирической оценкой, полученной по выборке, сэмплированной из данного распределения. Важным недостатком подобных методов является время их работы: чтобы получить хорошее приближение распределения, требуется сэмплировать много объектов, что может быть трудоёмко.

Чтобы обойти обозначенные недостатки описанных алгоритмов, а также тот факт, что каждый раз приходится перестраивать вероятностную модель и заново строить байесовский вывод, и в итоге получить простую, но гибкую и легко расширяемую модель для вероятностного тематического моделирования, был предложен подход Аддитивной Регуляризации Тематических Моделей (Additive Regularization For Topic Models, ARTM) [26–28]. Это многокритериальный подход, в котором к основному

критерию добавляется взвешенная сумма регуляризаторов. За счёт аддитивности оптимизация любых моделей и их комбинаций производится одним и тем же итерационным процессом. Для добавления регуляризатора в модель достаточно знать его частные производные по параметрам модели. Таким образом, ARTM — это не ещё одна тематическая модель, а общий подход к построению и комбинированию многих тематических моделей.

Однако, открытым остаётся вопрос о сходимости предложенного в рамках ARTM алгоритма. Известно, что на практике алгоритм сходится, тем не менее, теоретического обоснования алгоритма не было предложено. Итерации алгоритма ARTM можно проинтерпретировать как итерации Generalized Expectation Maximization (GEM) алгоритма [7] в случае наличия априорного распределения. Условия сходимости GEM алгоритмов хорошо изучены [33], их можно перенести их на алгоритм ARTM и получить ограничения на регуляризаторы.

1.2 Цель работы

Вопрос об условиях сходимости алгоритма ARTM является открытым, поэтому данная работа посвящена исследованию этой проблемы. Целями работы являются:

1. Изучение свойств сходимости алгоритма ARTM.
2. Поиск условий на регуляризаторы, способствующих сходимости. Требуется, чтобы данные условия легко проверялись для регуляризатора, а также, чтобы выполнялись на практике для применяемых регуляризаторов.
3. Анализ возможных модификаций алгоритма, улучшающих его сходимость.
4. Проведение эксперимента для проверки выполнения полученных условий, а также сравнения предложенных модификаций и стандартного алгоритма ARTM.

2 Аддитивная регуляризация тематических моделей

В данной главе будут введены используемые в работе обозначения, описаны основные алгоритмы, используемые при решении задачи вероятностного тематического моделирования, приведён краткий вывод алгоритма ARTM, а также будет проведено сравнение разных подходов с ARTM с целью показать преимуществ последнего.

2.1 Классическая тематическая модель

2.1.1 Постановка задачи и обозначения

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.

Введем 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, тема $t \in T$ является латентной (скрытой) переменной.

Требуется найти распределения терминов в темах $p(w|t) \equiv \varphi_{wt}$ для всех тем $t \in T$ и распределения тем в документах $p(t|d) \equiv \theta_{td}$ для всех документов $d \in D$. При этом делается ряд допущений.

Принимается гипотеза условной независимости $p(w|d, t) = p(w|t)$, и по формуле полной вероятности получается вероятностная модель порождения документа d:

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

Введем следующие обозначения:

$p_{tdw} \equiv p(t|d, w)$ — вероятность того, что появление термина w в документе d связано с темой t;

n_{dwt} — число троек (d, w, t) во всей коллекции. Другими словами, это число появлений термина w в связи с темой t в документе d;

$n_{dw} = \sum_{t \in T} n_{dwt}$ — число вхождений термина w в документ d, наблюдаемая величина;

$n_{td} = \sum_{w \in d} n_{dwt}$ — число вхождений терминов, связанных с темой t в документ d;

$n_{wt} = \sum_{d \in D} n_{dwt}$ — число появлений термина w в связи с темой t во всех документах коллеккции D;

$n_w = \sum_{d \in D} n_{dw}$ — число вхожений термина w в коллекцию;

$n_d = \sum_{t \in T} n_{td}$ — длина документа d;

$n_t = \sum_{w \in W} n_{wt}$ — "длина темы" t, то есть число появления в коллекции терминов, связанных с темой t;

$n = \sum_{d \in D} \sum_{w \in d} \sum_{t \in T} n_{dwt}$ — длина коллекции.

Правдоподобие — это плотность распределения выборки D:

$$p(D) = \prod_{i=1}^{n} p_i(d, w) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Рассмотрим вероятностную тематическую модель PLSA. Для этого вводится $p(D, \Phi, \Theta)$, где

$\Phi = (\varphi_{wt})_{W \times T}$ — искомая матрица терминов тем, $\varphi_{wt} \equiv p(w|t)$,

$\Theta = (\theta_{td})_{T \times D}$ — искомая матрица тем документов, $\theta_{td} \equiv p(t|d)$.

Запишем задачу максимизации правдоподобия:

$$p(D, \Phi, \Theta) = C \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}} = \prod_{d \in D} \prod_{w \in d} p(d|w)^{n_{dw}} C p(d)^{n_{dw}} \to \max_{\Phi, \Theta},$$

где C — нормировочный множитель, зависящий только от чисел n_{dw}. Прологарифмируем правдоподобие, получив задачу максимизации:

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

при ограничениях неотрицательности и нормировки

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

2.1.2 Итоговый алгоритм

Данная оптимизационная задача решается при помощи EM-алгоритма [12]:

E-шаг

На E-шаге, используя текущие значения параметров φ_{wt} и θ_{td}, по формуле Байеса вычисляются значения условных вероятностей:

$p_{tdw} \equiv p(t|d, w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}}$.

M-шаг

На M-шаге решается обратная задача: по условным вероятностям тем p_{tdw} вычисляются новые приближения φ_{wt} и θ_{td}. Можно заметить, что величина $n_{dwt} =$

$n_{dw}p(t|d,w) = n_{dw}p_{tdw}$ оценивает число вхождений термина w в документ d, связанных с темой t. При этом оценка не всегда является целым числом. Просуммировав n_{dwt} по документам d и по терминам w, получим оценки:

$n_{wt} = \sum_{d \in D} n_{dwt}$

$n_t = \sum_{w \in W} n_{wt}$

$n_{td} = \sum_{w \in d} n_{dwt}$

$n_d = \sum_{t \in T} n_{td}$

$\varphi_{wt} = \frac{n_{wt}}{n_t}$

$\theta_{td} = \frac{n_{dt}}{n_d}$

В исходном варианте алгоритм PLSA имеется ряд недостатков:

1. Медленная сходимость на больших коллекциях, так как матрицы Φ и Θ в базовом варианте обновляются после прохода всей коллекции, а также необходимость хранить трехмерную матрицу p_{tdw}. Эти проблемы могут быть решены принудительным более частым обновлениям Φ и Θ и внесением E-шага внутрь M-шага алгоритма.

2. Для алгоритма PLSA характерно переобучение, а также неединственность и неустойчивость решения. Это связано с большим числом параметров φ_{wt} и θ_{td} ($|W| \cdot |T| + |T| \cdot |D|$), на которые накладывается недостаточно ограничений. Кроме того, алгоритм PLSA неверно оценивает вероятность новых слов. Так, если $n_w = 0$, то $p(w|t) = 0$ для всех $t \in T$. Последний недостаток особо заметен в том случае, когда в контрольной выборке присутствует большое число новых терминов. Для устранения этой проблемы в алгоритм вводят регуляризации: сглаживание, разреживание, учёт дополнительной внешней информации.

3. Алгоритм не выделяет нетематические слова. В реальном тексте присутствуют термины, которые не относятся явно ни к одной из тем. Учет таких терминов возможен с помощью робастных тематических моделей, в которые добавляется шумовая и фоновая составляющие.

4. Алгоритм PLSA не позволяет управлять разреженностью. Действительно, если в начале работы алгоритма $\varphi_{wt} = 0$ или $\theta_{td} = 0$, то и после завершения работы алгоритма значения этих параметров останутся равными 0. Для борьбы с этим недостатком используют регуляризацию и постепенное разреживание.

2.2 Добавление регуляризатора

Для решения проблемы неединственности и неустойчивости вышеописанной оптимизационной задачи используется регуляризация. На искомое решение накладываются дополнительные ограничения. Подход ARTM [26–28] основан на идее многокритериальной регуляризации. Он позволяет строить модели, удовлетворяющие многим ограничениям одновременно. Каждое ограничение формализуется в виде регуляризатора — оптимизационного критерия $R_i(\Phi, \Theta) \to \max$, зависящего от параметров модели. Взвешенная сума всех таких критериев $R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$ максимизируется совместно с основным критерием правдоподобия.

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

при тех же ограничениях нормировки и неотрицательности.

Применение теоремы Каруша-Куна-Таккера позволяет получить систему уравнений для стационарных точек данной оптимизационной задачи. Решение данной системы методом простых итераций даёт EM-алгоритм со следующими формулами M-шага:

$$\begin{cases} \varphi_{wt} = \underset{w}{\mathsf{norm}} \left(n_{wt} + \varphi_{wt} \dfrac{\partial R}{\partial \varphi_{wt}} \right), \\ \theta_{td} = \underset{t}{\mathsf{norm}} \left(n_{td} + \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} \right). \end{cases}$$

где $\underset{x}{\mathsf{norm}}(y) = \frac{\max(y,0)}{\sum_x \max(y,0)}$. Операция $\max(y, 0)$ называется положительной срезкой.

Приведём примеры наиболее часто используемых регуляризаторов. Более подробное описание есть в статьях [26–28]:

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

Для комбинирования регуляризаторов 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.

2.3 Сравнение с другими подходами вероятностного тематического моделирования

Вероятностное тематическое моделирование развивается главным образом в рамках байесовского обучения и графических моделей. В байесовском подходе коллекция текстов описывается единой вероятностной порождающей моделью, при этом учёт дополнительных данных и формализация дополнительных ограничений производится через априорные распределения.
У такого подхода есть свои недостатки:

1. Не всякого рода знания удобно формализовать через априорные распределения. Попытка учесть больше знаний, чтобы построить более адекватную модель, приводит к значительному усложнению математического аппарата. В литературе почти нет работ по комбинированию тематических моделей, несмотря на их очевидную практическую востребованность.

2. Не факт, что естественный язык можно рассматривать как чисто статистическое явление. Одна из основных тенденций вычислительной лингвистики — создание гибридных моделей, объединяющих лучшие достижения статистических и лингвистических подходов. Лингвистические знания не всегда удобно описывать на вероятностном языке.

3. Большинство байесовских моделей вынужденно используют априорное распределение Дирихле [4]. Оно математически удобно благодаря сопряжённости с мультиномиальным распределением. Однако оно не моделирует каких-либо явлений естественного языка и не имеет убедительных лингвистических обоснований. Более того, оно противоречит естественному требованию разреженности, не допуская чистых нулей в матрицах Φ и Θ.

4. Априорное распределение Дирихле является слишком слабым регуляризатором. Проблему неустойчивости он не решает.

Учитывая эти проблемы, выделим преимущества подхода АРТМ:

1. В АРТМ регуляризаторы не обязаны быть априорными распределениями и иметь какую-либо вероятностную интерпретацию.

2. Регуляризатор Дирихле утрачивает свою особую роль, его не обязательно использовать в каждой модели для всех тем.

3. Математический аппарат очень прост: чтобы добавить регуляризатор, достаточно добавить его производные в формулы М-шага.

4. Многие байесовские тематические модели (или заложенные в них идеи) удаётся переформулировать через регуляризаторы.

5. Суммируя регуляризаторы, взятые из разных моделей, можно легко строить многоцелевые комбинированные модели.

3 Сходимость алгоритма ARTM

В данной главе будут изложены основные теоретические результаты работы. Будут доказаны теоремы о сходимости алгоритма ARTM, используя результаты о сходимостях GEM-алгоритмов. Поэтому первый раздел будет посвящён краткому описанию EM и GEM алгоритмов. После чего будут сформулированы и доказаны теоремы о сходимости. Последние разделы посвящены оценке изменения функционалов на M-шаге алгоритма ARTM, в которых были проведены теоретические оценки и были предложены возможные модификации для формулы M-шага, улучшающие оптимизацию.

3.1 Вероятностные EM и GEM алгоритмы

Модель PLSA уже была введена, правда рассматривалась $p(D, \Phi, \Theta)$. Предлагается задать модель иным способом, чтобы затем проделать вероятностный вывод EM алгоритма. Для этого расширим вероятностное пространство. С каждой словопозицией слова в документе d связывается одна определенная тема t. Неформально, это та тема, о которой думал автор, когда употребил конкретное слово w. Обозначим за Z 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:

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

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

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

и максимизируют данное выражение по Φ и Θ.

Задачу такого вида называют максимизацией неполного правдоподобия. Неполного — потому что из функции правдоподобия выведены скрытые переменные. Чтобы их вывести, производится суммирование по всем возможным значениям набора Z, которых может быть очень много. Поэтому просто использовать градиентные методы не получится. Здесь на помощь приходит мощный и очень полезный алгоритм машинного обучения — EM-алгоритм.

3.1.1 Общий подход: EM-алгоритм

Проведём вывод EM алгоритма в общем виде, поскольку появятся функционалы, которые будут использоваться в дальнейшем. Запишем задачу максимизации непол-

ного правдоподобия для некой вероятностной модели, в которой есть наблюдаемые переменные X, скрытые переменные Z и параметры Ω:

$$\log p(X|\Omega) \to \max_{\Omega}$$

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

$$\log p(X|\Omega) = \int q(Z) \log p(X|\Omega) dZ = \int q(Z) \frac{\log p(X,Z|\Omega)}{\log p(Z|X,\Omega)} dZ =$$

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

$$+ \underbrace{\int q(Z) \frac{q(Z)}{p(Z|X,\Omega)} dZ}_{KL(q(Z)\|p(Z|X,\Omega))} \quad (1)$$

Дивергенция Кульбака-Лейблера $KL(q(Z)\|p(Z|X,\Omega))$ 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 $\log p(X|\Omega)$. От максимизации $\log p(X|\Omega)$ 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. $F(q,\Omega) \to \max\limits_{\Omega}$

Максимизация $F(q,\Omega)$ по q эквивалентна минимизации $KL(q(Z)\|p(Z|X,\Omega))$, так как их сумма $\log p(X|\Omega)$ от q не зависит. Из свойств дивергенции Кульбака-Лейблера следует, что минимум, то есть 0, достигается при $q(Z) = p(Z|X,\Omega)$. То есть на первом шаге необходимо найти или оценить данное распределение.

Теперь рассмотрим второй шаг:

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

так как второе слагаемое не зависит от Ω, остаётся заметить, что была получена формула математического ожидания:

$$\int q(Z) \log p(X,Z|\Omega) dZ = \mathbf{E}_{q(Z)} \log p(X,Z|\Omega)$$

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.

E-step: $\underset{q(Z)}{\operatorname{argmin}} KL(q(Z)\|p(Z|X,\Omega)) = p(Z|X,\Omega)$

M-step: $\mathbf{E}_{q(Z)} \log p(X,Z|\Omega) \to \max_\Omega$

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.

3.1.2 EM-алгоритм максимизации неполного правдоподобия в модели PLSA

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

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

Чтобы найти эти вероятности, воспользуемся формулой Байеса:

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

Фактически это формула для p_{tdw} из главы 2.1.1. Теперь запишем выражение, которое нужно максимизировать на M-шаге:

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

Пронесём математическое ожидание внутрь суммы в силу независимости словопозиций, после чего распишем математическое ожидание по определению:

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

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

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

Таким образом, на каждом M-шаге нужно максимизировать данный функционал (в случае EM алгоритма) или строить увеличение (в случае GEM алгоритма).

3.1.3 Общий подход: введение априорного распределения

Пусть теперь стоит задача максимизации не апостериорной вероятности $p(X|\Omega)$, а максимизация полной вероятности $p(X,\Omega)$, учитывая некоторую априорную информацию о модели $p(\Omega)$. По формуле условной вероятности $p(X,\Omega) = p(X|\Omega)p(\Omega)$, повторяя старую декомпозицию(1), получаем оптимизационную задачу:

$$\log p(X|\Omega) + \log p(\Omega) = F(q,\Omega) + KL(q(Z)\|p(Z|X,\Omega)) + \log p(\Omega) \to \max_{\Omega}$$

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: $\mathbf{E}_{q(Z)} \log p(X,Z|\Omega) + \log p(\Omega) \to \max_{\Omega}$

Алгоритм ARTM имеет в точности такой вид, если интерпретировать τR 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.

3.2 ARTM как GEM алгоритм

Напомним, что в ARTM ставится задача максимизации следующего функционала:

$$L + \tau R = \sum_{w,d} n_{dw} \ln \sum_t \varphi_{wt}\theta_{td} + \tau R(\Phi,\Theta) \to \max$$

Как в GEM алгоритме, вводится дополнительный функционал:

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

где за p'_{tdw} обозначено $\frac{\varphi'_{wt}\theta'_{td}}{\sum\limits_t \varphi'_{wt}\theta'_{td}}$.

Требуется увеличивать значение данного функционала по Φ и Θ в сравнении с $Q(\Phi',\Theta',\Phi',\Theta')$ на каждой итерации. Запишем задачу максимизации данного функционала:

$$Q(\Phi,\Theta,\Phi',\Theta') \to \max_{\Phi,\Theta}.$$

Применив теорему Куна-Таккера, получим, что стационарная точка Q должна удовлетворять следующей системе:

$$\begin{cases} \varphi_{wt} = \underset{w}{\mathsf{norm}}\left(\sum\limits_d n_{dw} p'_{tdw} + \tau\varphi_{wt}\dfrac{\partial R}{\partial \varphi_{wt}}\right), \\ \theta_{td} = \underset{t}{\mathsf{norm}}\left(\sum\limits_w n_{dw} p'_{tdw} + \tau\theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right). \end{cases}$$

В итоге получаем систему уравнений, похожих на итерации ARTM. Это означает, что каждую итерацию ARTM можно интерпретировать как попытку приблизить решение максимизационной задачи функционала Q, итерируя систему уравнений для стационарной точки Q. В зависимости от того, какая точка будет браться начальной при итерировании системы уравнений, получаются разные варианты алгоритма ARTM.

Если начальное приближение — это $(\varphi_{wt}, \theta_{td})$, то получаем итерации:

$$\begin{cases} \varphi_{wt} = \underset{w}{\mathrm{norm}} \left(n_{wt} + \tau \varphi_{wt} \dfrac{\partial R}{\partial \varphi_{wt}} (\varphi_{wt}, \theta_{td}) \right), \\ \theta_{td} = \underset{t}{\mathrm{norm}} \left(n_{td} + \tau \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} (\varphi_{wt}, \theta_{td}) \right). \end{cases} \tag{2}$$

Эту формулу будем в дальнейшем называть стандартной формулой M-шага. Если же считать, что начальное приближение — это $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$, то:

$$\begin{cases} \varphi_{wt} = \underset{w}{\mathrm{norm}} \left(n_{wt} + \tau \dfrac{n_{wt}}{n_t} \dfrac{\partial R}{\partial \varphi_{wt}} \left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d} \right) \right), \\ \theta_{td} = \underset{t}{\mathrm{norm}} \left(n_{td} + \tau \dfrac{n_{td}}{n_d} \dfrac{\partial R}{\partial \theta_{td}} \left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d} \right) \right). \end{cases} \tag{3}$$

Эту формулу будем называть несмещённой модификацией M-шага.

Таким образом, интерпретируя ARTM как итерации GEM алгоритма, можно использовать результаты о сходимости GEM алгоритма.

3.3 Теоремы о сходимости ARTM

3.3.1 Итерации ARTM в обобщённом виде

Итерации ARTM можно записать в следующем виде:

E-step:

$p_{tdw} = \underset{t}{\mathrm{norm}}\, \varphi_{wt} \theta_{td}$

M-step:

$n_{wt} = \sum_d n_{dw} p_{tdw}, \quad n_{td} = \sum_w n_{dw} p_{tdw}$

$r_{wt} = \tau \varphi_{wt} \dfrac{\partial R}{\partial \varphi_{wt}}, \quad r_{td} = \tau \theta_{td} \dfrac{\partial R}{\partial \theta_{td}}$

$\varphi_{wt} = \underset{w}{\mathrm{norm}} \left(n_{wt} + r_{wt} \right), \quad \theta_{td} = \underset{t}{\mathrm{norm}} \left(n_{td} + r_{td} \right)$

Величины r_{wt} и r_{td} удобно называть регуляризационными добавками. Они являются функциями от Φ и Θ. Фактичеки, именно от их свойств зависит всё поведение алгоритма. Именно они и будут исследоваться в данной работе.

3.3.2 Ограниченные регуляризаторы

В работе Wu [33] были сформулированы достаточные условия для сходимости GEM алгоритма. Чтобы их сформулировать, нужно сначала ввести одно определение.

Определение 3.1 *Будем говорить, что* $A\colon X \to 2^X$ *— замкнутое point-to-set отображение, если из* $x_k \to x$, $x \in X$, $y_k \to y$ *и* $y_k \in A(x_k)$ *следует, что* $y \in A(x)$.

Итак, сформулируем теорему, немного изменив обозначения под ARTM:

Теорема 1 *Пусть* $\{\psi_p\}$ *— GEM последовательность, сгенерированная правилом* $\psi_{p+1} \in M(\psi_p)$, *где* M *— замкнутое point-to-set отображение. Пусть также значение* $L + \tau R$ *конечно и не уменьшается на итерациях, но при этом ограничено сверху,* $\|\psi_p - \psi_{p+1}\| \to 0$, *а множество стационарных точек* $L + \tau R$ *дискретно. Тогда* ψ_p *сходится к некоторой стационарной точке* $L + \tau R$.

Сходимость ARTM будет сведена к данной теореме, но для этого потребуется ввести новое определение.

Определение 3.2 *Будем говорить, что регуляризатор* τR *обладает свойством* δ*-регулярности, если на итерациях ARTM* $\forall t\ \exists w\colon n_{wt} + r_{wt} > \delta$ *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 данный регуляризатор сильно регулярен, при* $\delta = 0$ *будем говорить, что регуляризатор регулярен.*

Регулярность позволяет утверждать, что итерации ARTM корректно определены. Сильная же регулярность позволяет утверждать, что преобразования, которые производятся на итерациях, не только определены, но и непрерывны. Выполнение данного свойства на практике можно гарантировать следующим образом: если значение $n_{wt} + r_{wt}$ становится меньше δ, то вся тема зануляется и выкидывается, таким образом происходит селекция тем. Зануления тем будут происходить, если изначально задаётся заведомо большое число для их количества. Темы, содержащие слишком мало слов, удаляются, поскольку значение n_t получается небольшим.

Теорема 2 *Пусть* R *— ограниченная сверху и дифференцируемая функция, причем, как регуляризатор обладающая свойством регулярности. Также допустим, что значение* $Q(\Phi, \Theta, \Phi', \Theta')$ *конечно и не уменьшается в сравнении с* $Q(\Phi', \Theta', \Phi', \Theta')$ *на каждой итерации. Тогда при* d *и* w *таких, что* $n_{dw} > 0$ *выполнено*

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

где p^k_{tdw} — обозначение для значения величины на k-ой итерации.

Доказательство.

Заметим, что Q можно переписать следующим образом:

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

Пусть на итерации был переход в точку Φ'',Θ''. Q не уменьшается на итерациях, значит,

$$Q(\Phi'',\Theta'',\Phi',\Theta') \geqslant Q(\Phi',\Theta',\Phi',\Theta').$$

Подставим вместо Q его выражение:

$$L(\Phi'',\Theta'') + \tau R(\Phi'',\Theta'') + \sum_{d,w,t} n_{dw} p'_{tdw} \ln p''_{tdw} \geqslant L(\Phi',\Theta') + \tau R(\Phi',\Theta') + \sum_{d,w,t} n_{dw} p'_{tdw} \ln p'_{tdw}$$

$$\Delta(L+\tau R) \geqslant \sum_{d,w,t} n_{dw} p'_{tdw} \ln \frac{p'_{tdw}}{p''_{tdw}} = \sum_{d,w} n_{dw} KL(p'_{dw} || p''_{dw}) \geqslant 0.$$

Таким образом, $L+\tau R$ тоже не уменьшается. Но это ограниченная сверху функция, значит, $(L+\tau R)^k$ сходится при $k \to \infty$. Более того, при $n_{dw} > 0$:

$$KL(p^k_{tdw} || p^{k+1}_{tdw}) \leqslant \Delta(L+\tau R)^k \to 0.$$

■

Следствие 2.1 *Если в дополнение к условиям Теоремы 2 τR сильно регулярен, а r_{wt} и r_{td} непрерывны, то:*

$$|\varphi^k_{wt} - \varphi^{k+1}_{wt}| \to 0 \ \text{и} \ |\theta^k_{td} - \theta^{k+1}_{td}| \to 0$$

Доказательство. По неравеству Пинскера [24] $\|P-Q\|_1 \leqslant 2\sqrt{KL(P\|Q)}$. Поэтому сходимость по KL влечёт за собой сходимость по l_1 норме. Осталось заметить, что в потребованных условиях φ_{wt} и θ_{td} являются непрерывными функциями от p_{tdw}. А значит, сходимость вторых влечёт за собой сходимость первых. ■

Следствие 2.2 *В условия Следствия 1 все предельные точки последовательности (Φ^k,Θ^k) являются стационарными точками $L+\tau R$.*

Доказательство. Опишем коротко идею доказательства, более подробно и формально оно описано в [33]. Пусть Φ^0,Θ^0 — предельная точка последовательности (Φ^k,Θ^k). Известно, что выполнено:

$$Q(\Phi,\Theta,\Phi^0,\Theta^0) = L(\Phi,\Theta) + \tau R(\Phi,\Theta) + \sum_{d,w,t} n_{dw} p^0_{tdw} \ln p_{tdw}.$$

Поскольку φ^0, θ^0 — предельная точка, то значение Q уже нельзя увеличить, а, значит, производная по φ и по θ левой части равна нулю. При $\varphi = \varphi^0$ и $\theta = \theta^0$ KL достигает минимума, а, значит, и его производные равны нулю. Таким образом получается, что и производные $L + \tau R$ равны нулю, что и требовалось доказать. ■

Следствие 2.3 *Если в дополнение к условиям Следствия 1, множество стационарных точек $L+\tau R$ дискретно, то φ^k_{wt} и θ^k_{td} сходятся к стационарной точке $L+\tau R$.*

Доказательство.
Положим $M(\varphi, \theta) = \{artm(\varphi, \theta)\}$, где под $artm(\varphi, \theta)$ понимается применение формул ARTM 3.3.1. В условиях Следствия 1 $artm$ — непрерывное преобразование. Поэтому M — замкнутое point-to-set отображение. Остаётся заметить, что остальные условия Теоремы 1 тоже выполнены. ■

3.3.3 Неограниченные регуляризаторы

В предыдущем разделе была важна ограниченность R. Однако, в ARTM частно используется регуляризатор разреживания $-\alpha \sum_{w,t} \ln \varphi_{wt}$, который не является ограниченным. Тем не менее, на практике этот регуляризатор прекрасно работает. В данном разделе предлагается найти причину этого явления. Неформально, причиной может служить, что на практике есть машинная точность ε, и все значения, меньшие ε, считаются равными нулю. Это позволяет ограничить область значений φ_{wt} и θ_{td} снизу, и тем самым ограничить регуляризатор сверху. Также требуется разобраться с занулениями элементов матриц Φ и Θ на итерациях. При определённых ограничениях на регуляризатор эти зануления будут структурированными, что позволит провести анализ изменения функционала Q на итерациях.

Определение 3.3 *Будем говорить, что регуляризатор τR сохраняет 0, если на итерациях $n_{wt} = 0 \Rightarrow \varphi_{wt} = 0$ и $n_{td} = 0 \Rightarrow \theta_{td} = 0$.*

Это определение формализует следующие свойство итераций: 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 проверяется аналитически. На практике все регуляризаторы им обладают.

Определение 3.4 *Будем говорить, что регуляризатор τR ε-разреживающий, если на итерациях $\varphi_{wt}, \theta_{td} \notin (0, \varepsilon)$.*

Данное свойство позволяет формально учесть машинную точность (с этой точки зрения все регуляризаторы будут ε-разреживающими). Однако с точки зрения практики есть одна интересная особенность. Регуляризатор разреживания используется,

чтобы каждой теме принадлежало лишь небольшое число слов. Фактически, n_{wt} зануляется, если его значение меньше α, таким образом, после нормировки выполнено $\varphi_{wt} \geqslant \frac{n_{wt}-\alpha}{n_t} > 0$. На реальных коллекциях очень часто происходит следующее: характерные слова темы t имеют существенное значение n_{wt} (например, больше 1), а нехарактерные постепенно зануляются. В итоге получаем, что, начиная с некоторой итерации, $\varphi_{wt} \notin (0, \frac{1-\alpha}{n_t})$.

Есть альтернативный способ добиться данного ограничения: достаточно заменить регуляризатор разреживания на $-\alpha \ln \min(\varphi_{wt}, \alpha)$. В этом случае получим, что на M-шаге зануляются выражения меньше α и не изменяются остальные значения. В этом случае $\varphi_{wt} \notin (0, \frac{\alpha}{n_t})$.

Определение 3.5 *Будем говорить, что регуляризатор τR справедливый, если на итерациях $n_{dw} > 0 \Rightarrow \exists t: p_{tdw} > 0$.*

Это свойство — чистая формальность. Поскольку будут производиться разреживания, то необходимо случайно не занулить элемент матрицы $\Phi\Theta$ для которого $n_{dw} > 0$. Это привело бы к падению L до -∞. Данное свойство требует, чтобы такого не происходило. На практике оно обычно будет выполнено за счёт фоновых тем [27], поскольку они, как правило, дают небольшие вероятности для всех тем.

Итак, были введены три новых свойства регуляризатора, теперь можно доказать следующую теорему:

Теорема 3 *Пусть R — дифференцируемая функция при $\varphi_{wt}, \theta_{td} \in (0, 1]$, причем, как регуляризатор, сохраняющая 0, справедливая, ε-разреживающая и обладающая свойством регулярности. Также допустим, что значение $Q(\Phi, \Theta, \Phi', \Theta')$ конечно и не уменьшается в сравнении с $Q(\Phi', \Theta', \Phi', \Theta')$, начиная с некоторой итерации. Тогда выполнено:*

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

Доказательство.

Поскольку регуляризатор сохраняет 0, то с некоторой итерации множество позиций с нулевыми значениями в матрице Φ и Θ стабилизируется и не будет больше изменяться. Это очевидно следует из того факта, что множество всех позиций конечно. Обозначим стаблизиризовавшееся множество ненулевых позиций за Ω. В силу того, что регуляризатор ε-разреживающий, значения Φ и Θ в позициях из Ω будут $\geqslant \varepsilon$. Но R — дифференцируемая функция при $\varphi_{wt}, \theta_{td} \in [\varepsilon, 1]$, а, значит, непрерывная и ограниченная. Далее можно повторить рассуждения Теоремы 2, ограничивших значениями φ_{wt} и θ_{td} только в этих позициях. ■

Так же как и в случае Теоремы 2, данная теорема будет иметь аналогичные три следствия. Они не будут дублироваться здесь, приведём только итоговую теорему, объединяющую все утверждения.

Теорема 4 *Пусть R — дифференцируемая функция при $\varphi_{wt}, \theta_{td} \in (0,1]$, причем, как регуляризатор, сохраняющая 0, справедливая, ε-разреживающая и обладающая свойством сильной регулярности, а r_{wt} и r_{td} непрерывны. Также допустим, что значение $Q(\Phi, \Theta, \Phi', \Theta')$ конечно и не уменьшается в сравнении с $Q(\Phi', \Theta', \Phi', \Theta')$, начиная с некоторой итерации. Тогда, если множество стационарных точек $L+\tau R$ дискретно при любой фиксации множества ненулевых элементов в матрицах Φ и Θ, то φ_{wt}^k и θ_{td}^k сходятся к стационарной точке $L+\tau R$ при ограничении на некоторое множество ненулевых позиций в матрицах Φ и Θ.*

Проанализируем регуляризатор разреживания с точки зрения данной теоремы в случае стандартных формул М-шага ($r_{wt} = \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}$). Свойство ε-разреживания было обсуждено при определении данного свойства. Свойство справедливости обычно выполняется за счёт фоновых тем [27] . Сохранение нуля данным регуляризатором и непрерывность r_{wt} очевидны. Единственный важный момент — это конечность Q на итерациях. Если произойдёт зануление элементов матриц Φ и Θ, то значение регуляризатора уходит в бесконечность. Чтобы избежать этого эффекта, надо считать $R = -\alpha \sum\limits_{w:\ \varphi_{wt}>0} \ln \varphi_{wt}$.

Теперь можно описать, что происходит на итерациях ARTM (в предположении увеличения Q). Итерации можно разбить на два этапа: селекция ненулевых позиций и оптимизация. На первом этапе при помощи регуляризатора выбирается множество ненулевых позиций итогового решения. Понятно, что параллельно ведётся и оптимизация $L + \tau R$, но из-за наличия положительной срезки в нормировке на М-шаге этот этап очень сложно анализировать. Его стоит воспринимать как подготовка начального приближения. На втором этапе оптимизация выходит на первый план. В силу того, что множество нулевых позиций не изменяется, положительную срезку в формулах можно убрать. Это облегчает анализ, который будет проведён в соответствующей главе.

3.4 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 доста-

точных условий сходимости. Сходимость итерационного процесса неразрывно связана со свойстами его траектории. Поэтому удалось сопоставить свойства траектории процесса с изменениями $L + \tau R$.

Теорема 5 *Пусть выполнены условия теоремы 2. Тогда сходимость ряда*

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

влечёт за собой сходимость ряда

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

Доказательство.

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

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

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

Следствие 5.1 *В условиях теоремы 2 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

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

А сходимость данной последовательности уже была доказана. ■

Следствие 5.2 *В условиях теоремы 4 условие дискретности множества стационарных точек можно заменить условием сходимости ряда*

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

К сожалению, это абсолютно 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.

3.5 Изменение регуляризированного правдоподобия на итерациях ARTM

Важным условием сходимости алгоритма ARTM является неуменьшение значения Q на M-шаге. В данной главе изучается поведение алгоритма на M-шаге и проводятся оценки изменения функционалов L, R и Q. На основании результатов данного анализа будет предложена модификация M-шага (в дополнение к замене φ_{wt} и θ_{td} несмещёнными оценками).

В данной главе под Q' будем понимать $\sum\limits_{d,w,t} n_{dw} p'_{tdw} \ln \varphi_{wt}\theta_{td}$. То есть $Q = Q' + \tau R$.

3.5.1 Общий анализ

Напомним, имеется два набора формул для регуляризационных добавок на M-шаге:

$$\begin{cases} r_{wt} = \tau\varphi_{wt}\dfrac{\partial R}{\partial \varphi_{wt}}, \\ r_{td} = \tau\theta_{td}\dfrac{\partial R}{\partial \theta_{td}}, \end{cases} \quad \text{и} \quad \begin{cases} r_{wt} = \tau\dfrac{n_{wt}}{n_t}\dfrac{\partial R}{\partial \varphi_{wt}}\left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d}\right), \\ r_{td} = \tau\dfrac{n_{td}}{n_d}\dfrac{\partial R}{\partial \theta_{td}}\left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d}\right). \end{cases}$$

Обновление параметров происходит по формулам:

$$\begin{cases} \varphi_{wt} = \underset{w}{\mathsf{norm}}\Big(\sum\limits_d n_{dw} p_{tdw} + r_{wt}\Big), \\ \theta_{td} = \underset{t}{\mathsf{norm}}\Big(\sum\limits_w n_{dw} p_{tdw} + r_{td}\Big). \end{cases}$$

Провести анализ суммарного изменения функционала Q по таким формулам не представляется возможнным. Поэтому предлагается разложить это преобразование на два этапа. Первый этап — максимизация Q':

$$\begin{cases} \varphi_{wt} = \underset{w}{\mathsf{norm}}\, n_{wt}, \\ \theta_{td} = \underset{t}{\mathsf{norm}}\, n_{td}. \end{cases}$$

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

$$\begin{cases} \varphi_{wt} = \underset{w}{\mathsf{norm}}\Big(n_{wt} + r_{wt}\Big), \\ \theta_{td} = \underset{t}{\mathsf{norm}}\Big(n_{td} + r_{td}\Big). \end{cases}$$

Таким образом, изменения функционалов будут оцениваться отдельно на каждом этапе. На первом происходит переход в точку $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$, максимизируя Q', а на втором проводится максимизация R.

Как говорилось в главе 3.2, есть две естественные формулы для регуляризационных добавок. Напомним их на примере r_{wt}. В первом случае $r_{wt} = \tau\varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\left(\varphi_{wt}, \theta_{td}\right)$,

а во втором $r_{wt} = \frac{n_{wt}}{n_t}\frac{\partial R}{\partial \varphi_{wt}}\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$. Первый способ предлагает определить добавку в точке $(\varphi_{wt}, \theta_{td})$ с предыдущей итерации, второй способ определяет добавку уже в новой точке $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$. Интуитивно понятно, что второй способ должен лучше максимизировать R, так как он строит добавки по более актуальной информации.

Для первого варианта не удалось провести анализ, так как неясно, как связаны градиенты R в точках $(\varphi_{wt}, \theta_{td})$ и $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$. Для второго варианта анализ оказался успешнее.

Введём новую функцию R', зависящую от n_{wt} следующим образом: $R'(n_{wt}, n_{td}) = R\left(\frac{n_{wt}}{\sum\limits_w n_{wt}}, \frac{n_{td}}{\sum\limits_t n_{td}}\right)$. Также введём обозначение $g_{wt} \equiv \frac{\partial R'}{\partial n_{wt}}$ и $g_{td} \equiv \frac{\partial R'}{\partial n_{td}}$. Обратим внимание, что это функции. Для них справедливо следующее утверждение:

Утверждение 3.5.1 *Для g_{wt} и g_{td} выполнено*

$$\begin{cases} g_{wt} = \dfrac{1}{n_t}\sum\limits_u \left(\dfrac{\partial R}{\partial \varphi_{wt}} - \dfrac{\partial R}{\partial \varphi_{ut}}\right)\varphi_{ut}, \\ g_{td} = \dfrac{1}{n_d}\sum\limits_s \left(\dfrac{\partial R}{\partial \theta_{td}} - \dfrac{\partial R}{\partial \theta_{sd}}\right)\theta_{sd}. \end{cases} \tag{4}$$

Доказательство.

$\varphi_{wt} = \frac{n_{wt}}{\sum_w n_{wt}}$, поэтому

$$\frac{\partial \varphi_{ut}}{\partial n_{wt}} = \frac{\partial \frac{n_{ut}}{\sum\limits_v n_{vt}}}{\partial n_{wt}} = \frac{\frac{\partial n_{ut}}{\partial n_{wt}}}{\sum\limits_v n_{vt}} - \frac{n_{ut}}{(\sum\limits_v n_{vt})^2} = I\{u = w\}\frac{1}{n_t} - \frac{\varphi_{ut}}{n_t} = \frac{1}{n_t}\Big(I\{u = w\} - \varphi_{ut}\Big).$$

А, значит,

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

Для $\frac{\partial R'}{\partial n_{td}}$ формула доказывается аналогично. ■

Теперь докажем следующую лемму:

Лемма 3.5.2 *Для несмещённой модификации M-шага в ходе регуляризационного преобразования без занулений угол между вектором изменений и градиентом R острый, если градиент ненулевой и $\tau > 0$.*

Доказательство.

При регуляризационном преобразовании без занулений $\Delta n_{wt} = \tau \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}$, а для несмещённой модификации $\varphi_{wt} = \frac{n_{wt}}{\sum\limits_w n_{wt}}$. Отсюда

$$\langle \Delta n, \nabla R'(n_{wt}, n_{td})\rangle = \sum_{w,t,u} \frac{1}{n_t}\left(\frac{\partial R}{\partial \varphi_{wt}} - \frac{\partial R}{\partial \varphi_{ut}}\right)\tau \frac{\partial R}{\partial \varphi_{wt}}\varphi_{wt}\varphi_{ut}.$$

Если переобозначить u за w и наоборот, то

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

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

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

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

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

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

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

Значит, градиент нулевой — противоречие. Поэтому неравенство строгое и угол острый. ■

В главе 3.3.3 было показано, что при определённых условиях (Теорема 3) на регуляризатор занулений не будет, начиная с некоторой итерации. Таким образом, если коэффициент τ не слишком большой, то изменение n_{wt} будет небольшим. Поэтому при регуляризационном преобразовании будет происходить увеличение R в силу локального изменения вдоль градиента.

При $\tau < 0$ острым будет угол с антиградиентом R, что означает локальное уменьшение R.

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

Остаётся понять, как изменяется данный функционал на первом шаге. Начиная с некоторого момента, изменения параметров становятся небольшими, а это означает, что максимизируется значение $Q' + \tau R$ в локальной окрестности точки $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$, в которой находится и исходная точка $(\varphi_{wt}, \theta_{td})$, то есть она была потенциальным кандидатом при выборе улучшения. Поэтому, если выбирается другое направление, то происходит увеличение значения $Q' + \tau R$ по сравнению с значением в $(\varphi_{wt}, \theta_{td})$.

3.5.2 Использование градиента регуляризатора

У вышеописанного рассуждения есть два допущения. Первое: считается, что изменения φ и θ невелики. Обычно так и есть после нескольких первых итераций, когда приближённые оценки параметров будут найдены и большие скачки в пространстве матриц перестают происходить. Второе: считается, что происходит локальная максимизация $Q' + \tau R$, однако, было доказано, что происходит увеличение, а не максимизация. Тем не менее, существует несколько способов обойти это условие.

Во-первых, на практике угол между вектором регуляризационных поправок и градиентом регуляризатора весьма острый, что позволяет производить требуемое увеличение. Во-вторых, можно выбирать направление между предлагаемым и направлением на старую точку. В третьих, можно использовать значения g_{wt} и g_{td} в качестве регуляризационных добавок. Тогда направление изменения при регуляризационном преобразовании будет совпадать с направлением градиента, а, значит, будет локальная максимизация. То есть предлагается использовать следующую формулу для регуляризационных добавок на М-шаге:

$$\begin{cases} r_{wt} = \tau A_t g_{wt}\left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d}\right) = \tau A_t \dfrac{1}{n_t}\left[\dfrac{\partial R}{\partial \varphi_{wt}} - \sum\limits_u \varphi_{ut}\dfrac{\partial R}{\partial \varphi_{ut}}\right]\left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d}\right), \\ r_{td} = \tau B_d g_{td}\left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d}\right) = \tau B_d \dfrac{1}{n_d}\left[\dfrac{\partial R}{\partial \theta_{td}} - \sum\limits_s \theta_{sd}\dfrac{\partial R}{\partial \theta_{sd}}\right]\left(\dfrac{n_{wt}}{n_t}, \dfrac{n_{td}}{n_d}\right), \end{cases} \tag{5}$$

где A_t и B_d — это некоторые константы, характеризующие величину шага вдоль градиента по данной теме (документу). Для экспериментов использовались два простых варианта: $A_t = B_d = 1$ и $A_t = B_d = 50$.

Такие добавки можно эффективно вычислить, поскольку второе слагаемое общее для всех слов (тем). Поэтому его можно считать кумулятивно при первом проходе, на котором считается первое слагаемое, а затем на втором проходе вычесть его из всех добавок. Таким образом, асимптотика времени работы алгоритма не увеличится, а константа в асимптотике незначительно возрастёт.

Градиент регуляризатора можно использовать ещё одним способом, сделав подмену регуляризатора. Пусть регуляризатор S такой, что

$$\varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} = C_t g_{wt}(\varphi_{wt}, \theta_{td}), \tag{6}$$

где C_t — некоторая константа, зависящая только от темы. Тогда можно применить рассуждения теорем о сходимости к регуляризатору R, но для анализа использовать функционал $Q' + \tau S$. Тогда направление изменения при регуляризационном преобразовании для несмещённой модификации (3) будет совпадать с направлением градиента, а, значит, будет локальная максимизация. К сожалению, решение данной

системы уравнений в частных производных затруднено тем фактом, что матрица коэффициентов вырождена, что означает, что решение существует не всегда (на текущий момент есть гипотеза, что необходимым и достаточным условием является $\sum \varphi_{wt}^2 \frac{\partial R}{\partial \varphi_{wt}} = 0$). Анализ данного уравнения является перпективой дальнейшего исследования.

3.5.3 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 ешё одним важным свойством: их воздействие можно считать отдельно. Поясним, что именно имеется в виду. Пусть $R = R_1 + R_2$, где R_1 — аналитический регуляризатор. На М-шаге необходимо построить увеличение функционала $Q + \tau R = Q + \tau R_1 + \tau R_2$. По формулам М-шага вычисляется $(n_{wt},\ n_{td})$ как точка максимума Q, а затем увеличивается R. Однако, можно определить (n_{wt}, n_{td}) как точку максимума $Q + \tau R_1$ (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 $Q' + \tau R$ тоже вогнутая функция, и, следовательно, имеет единственный максимум. При некоторых дополнительных допущениях будет происходить увеличение $Q' + \tau R$. Однако, в случае вогнутого регулязиратора 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 некорректо поставленной, поскольку максимум оптимизируемой функции равен бесконечности. Более подробно данная проблема рассматривалась в главе 3.3.3.

Произвольные регуляризаторы. Для произвольных регуляризаторов было доказано увеличение R при регуляризационном преобразовании. При дополнитель-

ных условиях оно преобразуется в увеличение $Q' + \tau R$ на итерациях. Здесь наиболее интересно научиться делать подмену регуляризатора по системе уравнений(6).

3.5.4 Различия предложенных модификаций М-шага

Пусть $R = -\tau \sum_{w,t} \varphi_{wt}$. Формально, он не должен влиять на оптимизацию, поскольку равен константе при ограничениях задачи. Однако, стандартные формулы дадут следующий М-шаг:

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

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:

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

где $\underset{\tau}{\mathsf{sparse}}(x, y) = x$, 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. Использование градиентной добавки даёт

$$r_{wt} \propto \frac{\partial R}{\partial \varphi_{wt}} - \sum_u \varphi_{ut} \frac{\partial R}{\partial \varphi_{ut}} = -\tau + \tau = 0,$$

то есть в точности PLSA.

Этот пример показывает, что градиентная добавка менее склонна к занулению параметров. Это наводит на мысль, что вначале можно использовать обычные формулы М-шага, или несмещённые оценки для селекции тем. А затем использовать градиентную поправку для более точной настройки параметров.

3.6 Стремление коэффициентов к нулю

Основным инструментом для анализа регуляризационного преобразования был подсчёт градиента не по φ_{wt}, а по n_{wt}. Используя данный подход, докажем следующее утверждения для коэффициентов регуляризации, стремящихся к нулю.

Утверждение 3.6.1 *Существует такая константа γ, что если $\tau_n \leqslant \gamma \Delta Q'_n$, а также $\frac{1}{n_t}\frac{\partial R}{\partial \varphi_{wt}}(n_{wt}, n_{td})$ — 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 $\Delta Q'_n \geqslant 0$.*

Доказательство.

Для лаконичности рассмотрим случай $R(\Phi, \Theta) = R(\Phi)$.

При регуляризационном сглаживании $\Delta n_{wt} = \left(n_{wt} + \tau_n \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right)_+ - n_{wt}$.

$$\left(n_{wt} + \tau_n \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right)_+ - n_{wt} \leqslant n_{wt} + \left|\tau_n \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right| - n_{wt} \leqslant \left|\tau_n \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right|$$

$$\left(n_{wt} + \tau_n \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right)_+ - n_{wt} \geqslant n_{wt} - \left|\tau_n \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right| - n_{wt} \geqslant -\left|\tau_n \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right|$$

$$|\Delta n_{wt}| \leqslant \tau_n \left|\varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}}\right| \leqslant \tau_n n_{wt} C \leqslant \tau_n n_w C$$

Подставим значение градиента R (4) :

$$\left|\langle \Delta n, \nabla R(n_{wt}, n_{td}) \rangle\right| = \left|\sum_{w,t,u} \frac{1}{n_t} \left(\frac{\partial R}{\partial \varphi_{wt}} - \frac{\partial R}{\partial \varphi_{ut}}\right) \Delta n_{wt} \varphi_{ut}\right| \leqslant 2C^2 \sum_w n_w \tau_n \leqslant (2\gamma C^2 \sum_w n_w) \Delta Q'_n$$

Если $2\gamma C^2 \sum_w n_w < 1$, то изменение Q'_n при регуляризационном преобразовании меньше чем при максимизации Q'_n, а, значит, суммарный эффект будет положительным.

■

4 Практические исследования

Данная глава посвящена экспериментам, проведённым с целью проверки выполнения полученных условий сходимости, а также сравнения двух предложенных модификаций (несмещённой и градиентной) и стандартного алгоритма ARTM. В конце главы будут сделаны выводы о полученных результатах.

4.1 Используемая коллекция данных

Для экспериментов была нужна достаточно большая коллекция документов, с заранее известным и небольшим числом тем. Стандартная для тематического моделирования коллекция 20Newsgroups не подошла, так как там слишком мало документов. Википедия не подошла, поскольку содержит слишком много тем. Поэтому была собрана собственная коллекция документов. Были скачены статьи со спортивного сайта sports.ru по разным видам спорта, темой документа считался вид спорта. Получившаяся коллекция документов состоит из 7 спортивных направлений примерно по 3000 статей в каждом. Также была проведена лемматизация текста и удаление стоп слов. Итоговые параметры коллекции: $|W| = 18831,\ |D| = 21001,\ |T| = 7$.

4.2 Сравниваемые алгоритмы

Было проведено сравнение трёх вариантов М-шага: стандартного варианта, несмещённой модификации М-шага и градиентной модификации М-шага. Приведём описание используемых алгоритмов.

Алгоритм 1: ARTM. Стандартный М-шаг

procedure MStep(n_{wt}, n_{td}, φ_{wt}, θ_{td})

1. Для всех пар w, t вычислить $r_{wt} = \tau\varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}(\varphi_{wt}, \theta_{td})$
2. Для всех пар t, d вычислить $r_{td} = \tau\theta_{td}\frac{\partial R}{\partial\theta_{td}}(\varphi_{wt}, \theta_{td})$
3. Для всех пар w, t вычислить $p_{wt} = \max(n_{wt} + r_{wt}, 0)$
4. Для всех пар t, d вычислить $p_{td} = \max(n_{td} + r_{td}, 0)$
5. Вычислить нормировочные множители $p_t = \sum_w p_{wt}$ и $p_d = \sum_t p_{td}$
6. Для всех пар w, t вычислить $\varphi_{wt} = \frac{p_{wt}}{p_t}$
7. Для всех пар t, d вычислить $\theta_{td} = \frac{p_{td}}{p_d}$
8. Вернуть матрицы φ_{wt}, θ_{td}

Алгоритм 2: ARTM. Несмещённый M-шаг

procedure MStep(n_{wt}, n_{td}, φ_{wt}, θ_{td})

1. Вычислить нормировочные множители $n'_t = \sum_w n_{wt}$ и $n'_d = \sum_t n_{td}$
2. Для всех пар w, t вычислить $r_{wt} = \tau \frac{n_{wt}}{n'_t} \frac{\partial R}{\partial \varphi_{wt}}\left(\frac{n_{wt}}{n'_t}, \frac{n_{td}}{n'_d}\right)$
3. Для всех пар t, d вычислить $r_{td} = \tau \frac{n_{td}}{n'_d} \frac{\partial R}{\partial \theta_{td}}\left(\frac{n_{wt}}{n'_t}, \frac{n_{td}}{n'_d}\right)$
4. Для всех пар w, t вычислить $p_{wt} = \max(n_{wt} + r_{wt}, 0)$
5. Для всех пар t, d вычислить $p_{td} = \max(n_{td} + r_{td}, 0)$
6. Вычислить нормировочные множители $p_t = \sum_w p_{wt}$ и $p_d = \sum_t p_{td}$
7. Для всех пар w, t вычислить $\varphi_{wt} = \frac{p_{wt}}{p_t}$
8. Для всех пар t, d вычислить $\theta_{td} = \frac{p_{td}}{p_d}$
9. Вернуть матрицы φ_{wt}, θ_{td}

Алгоритм 3: ARTM. Градиентный M-шаг

procedure MStep(n_{wt}, n_{td}, φ_{wt}, θ_{td})

1. Вычислить нормировочные множители $n'_t = \sum_w n_{wt}$ и $n'_d = \sum_t n_{td}$
2. Для всех пар w, t вычислить $r_{wt} = \tau \frac{1}{n'_t} \frac{\partial R}{\partial \varphi_{wt}}\left(\frac{n_{wt}}{n'_t}, \frac{n_{td}}{n'_d}\right)$
3. Для всех пар t, d вычислить $r_{td} = \tau \frac{1}{n'_d} \frac{\partial R}{\partial \theta_{td}}\left(\frac{n_{wt}}{n'_t}, \frac{n_{td}}{n'_d}\right)$
6. Вычислить $r_t = \sum_w \frac{n_{wt}}{n'_t} r_{wt}$ и $r_d = \sum_t \frac{n_{td}}{n'_d} r_{td}$
4. Для всех пар w, t вычислить $p_{wt} = \max(n_{wt} + r_{wt} - r_t, 0)$
5. Для всех пар t, d вычислить $p_{td} = \max(n_{td} + r_{td} - r_t, 0)$
6. Вычислить нормировочные множители $p_t = \sum_w p_{wt}$ и $p_d = \sum_t p_{td}$
7. Для всех пар w, t вычислить $\varphi_{wt} = \frac{p_{wt}}{p_t}$
8. Для всех пар t, d вычислить $\theta_{td} = \frac{p_{td}}{p_d}$
9. Вернуть матрицы φ_{wt}, θ_{td}

Алгоритм 4: ARTM. E-шаг

procedure EStep($\varphi_{wt}, \theta_{td}$)

1. Инициализировать переменные n_{wt} и n_{td}
2. Для каждого документа d проделать строки 3-6
3. Вычислить матрицу $p_{tdw} = \varphi_{wt}\theta_{td}$
4. Отнормировать величины p_{tdw} по t
5. Для каждого слова w в документе d увеличить n_{wt} на $n_{dw}p_{tdw}$
6. Для каждого слова w в документе d увеличить n_{td} на $n_{dw}p_{tdw}$
7. Вернуть матрицы n_{wt}, n_{td}

Алгоритм 5: ARTM. EM-алгоритм

procedure EM(iters)

1. Инициализировать начальное приближение $\varphi_{wt}, \theta_{td}$
2. Повторить iters раз строки 3-4
3. $n_{wt},\ n_{td}\ =\ EStep(\varphi_{wt}, \theta_{td})$
4. $\varphi_{wt},\ \theta_{td}\ =\ MStep(n_{wt}, n_{td}, \varphi_{wt}, \theta_{td})$
5. Вернуть матрицы φ_{wt}, θ_{td}

4.3 Исследуемые величины

В качестве регуляризатора в экспериментах был выбран регуляризатор декорреляции: $R = \sum_w \sum_{s \neq t} \varphi_{wt}\varphi_{ws}$. Этот регуляризатор был выбран как самый простой из неаналитических и невыпуклых регуляризаторов. Использовались четыре разных значения τ: -10^5, -10^6, -10^7 и -10^8. Значения отрицательные, поскольку нужно уменьшать корреляции тем. Также проверялись разные количества тем: 3, 10, 30, чтобы проверить поведение алгоритма в случаях недооценки, достаточно точной оценки и переоценки количества тем. Измерялись следующие величины:

1. Минимальное ненулевое значение в матрицах Φ и Θ. Таким образом, будет проверено выполнение свойства отделимости от нуля.

2. Значение L, R и $L + \tau R$ на итерациях. Это показатель качества оптимизации качества оптимизации.

3. Изменение R при выполнении М-шага. Тут есть небольшая особенность. Дело в том, что значения r_{wt} по модулю при градиентной модификации на порядок меньше значений при стандартном М-шаге или несмещённой модификации. Поэтому, чтобы более точно сравнить качество оптимизации, было решено нормировать изменение R на норму (использовались l_1 и l_2 нормы) вектора r_{wt}, проверяя таким способом качество выбранного направления. Помимо использования g_{wt} как r_{wt} было попробовано использовать g_{wt} умноженные на 50. Эта константа была выбрана из соображений, чтобы изменения R были по порядку такие же, как и у других алгоритмов.

4. Минимальные ненулевые значения n_t на М-шаге.

Алгоритм запускался из 20 начальных приближений, одинаковых для всех запусков, после чего сравнивались средние значения целевых метрик. При фиксированных параметрах расчёт 20 начальных приближений работает полчаса.

4.4 Особенности реализации

Во-первых, требовалось обеспечить, что φ и θ отделимы от нуля. Однако, было замечено, что в матрицах Φ и Θ некоторые элементы стремятся к нулю, но обнуляются посредством машинной точности только спустя очень большое количество итераций. Поэтому при выполнении М-шага производилось очень слабое разрежение. То есть n_{wt} и n_{td} меньшие 10^{-6} зануялись. Формально, такое воздействие можно задать регуляризатором, об этом говорилось в главе 3.3.3. Значения данного регуляризатора очень малы и не влияют на значение замеряемых функционалов.

Во-вторых, из-за ошибок округления возникали граничные эффекты при вычислении логарифма правдоподобия, поэтому при вычислении логарифмов в формулах использовался $\log(x+\varepsilon)$, где ε — маленькая константна равная 10^{-20}. Благодаря этому логарифм правдоподобия не устремлялся к минус бесконечности в плохих точках, но становился достаточно малым, чтобы заметить этот эффект. Фактически, размером этой константы определяется, в окрестности какого числа будут получаться значения логарифма правдоподобия.

4.5 Результаты и выводы

В данной главе будут приведены полученные измерения целевых функционалов, сделаны технические выводы на их основании, а также подведены итоги экспериментов.

4.5.1 Значения $L + \tau R$

ARTM выводится как оптимизация функции $L + \tau R$, поэтому в первую очередь важно поведение именно данного функционала на итерациях.

Рис. 1: $|T| = 3$. В основном градиентные поправки показывают немного лучший результат чем стандартный и несмещённый M-шаги.

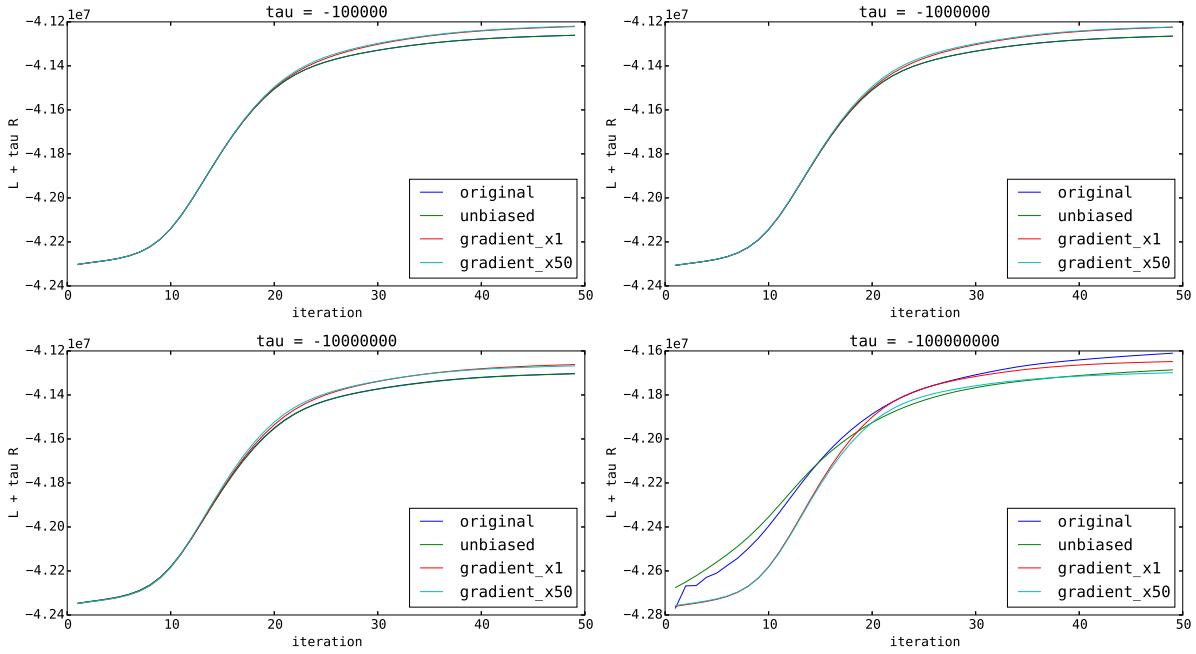


Рис. 2: $|T| = 10$. Градиентные поправки немного хуже остальных, но при большом $|\tau|$ показывают ощутимо лучший результат.

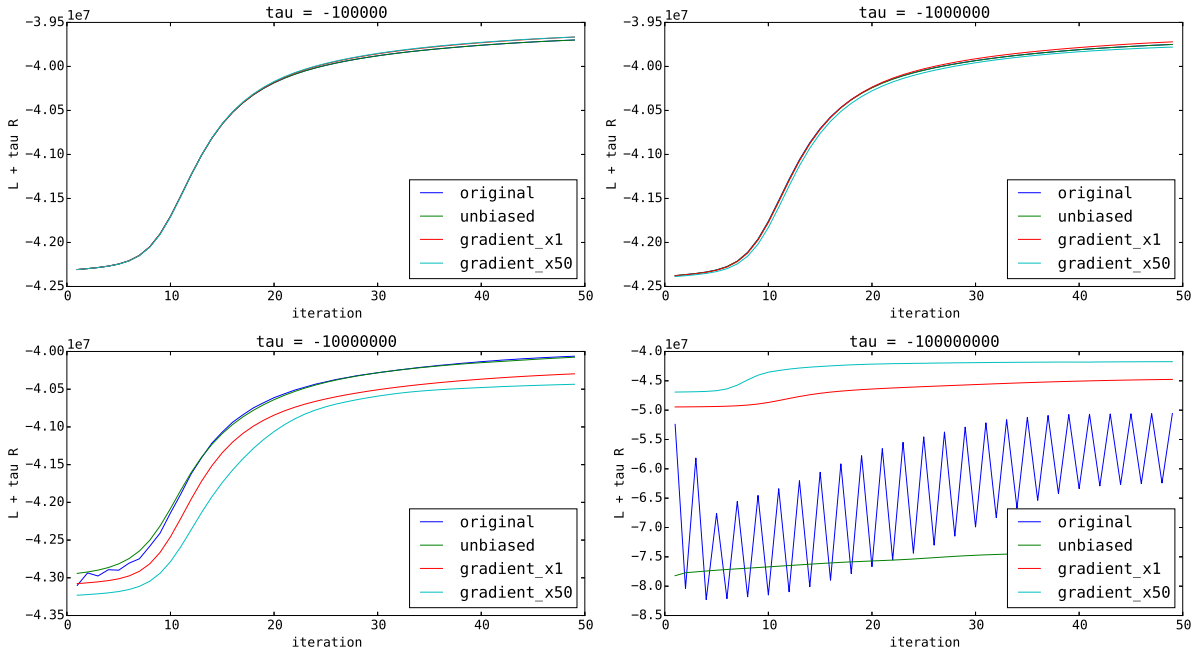
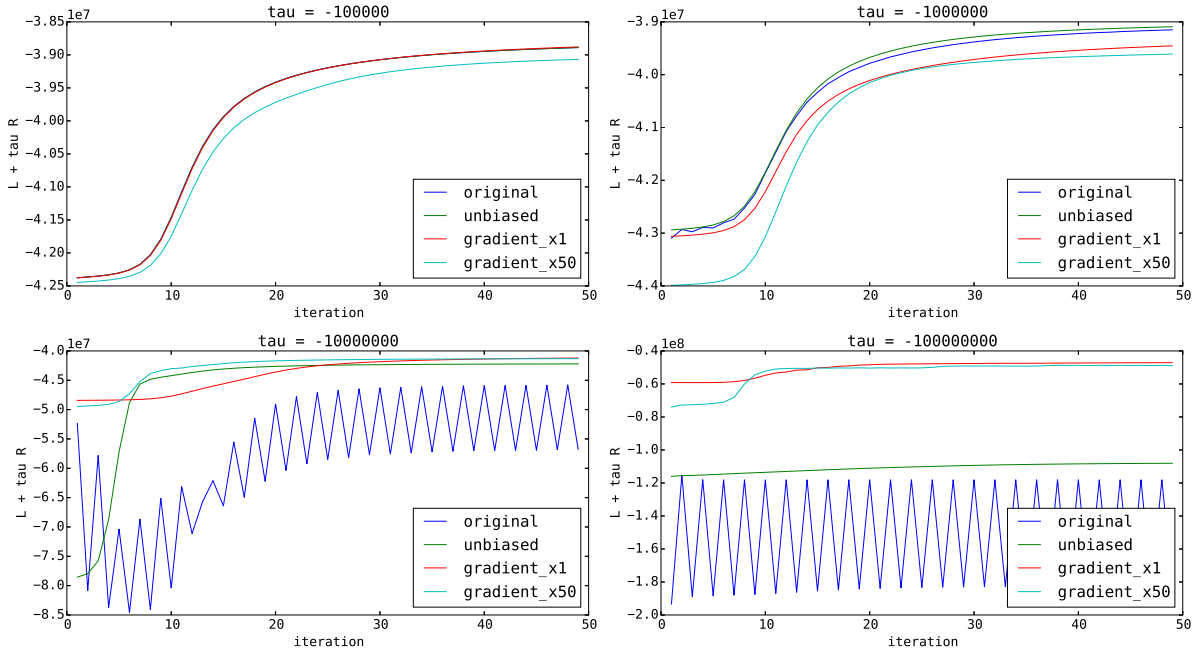


Рис. 3: $|T| = 30$. Скачки начинаются при меньших $|\tau|$, до этого градиентные поправки так же немного хуже.



Иногда градиентная поправка показывает чуть более слабые результаты. Это можно объяснить тем, что константы, на которые домножается градиент регуляризатора, выбраны не совсем удачно. Подбор данных констант — это направление будущих исследований.

Также стоит отметить, что графики несмещённой модификации и стандартного М-шага практически не отличаются. Однако, с ростом $|\tau|$ можно увидеть, что начинаются скачки на графике стандартной формулы, а на графике несмещённой модификации их нет. Это связано с тем, как именно считаются регуляризационные поправки. Подробнее этот эффект будет обсуждён в следующей главе на графиках значений R, где он будет сильнее заметен.

Зависимость от числа тем удобнее рассмотреть на значениях функционала R.

При очень большом значении τ графики для стандартной и несмещённой оценки резко падают. Этот эффект вызван следующим: стандартные формулы и несмещённая модификация зануляют слишком много параметров, что приводит к падению правдоподобия (нарушается свойство справедливости регуляризатора 3.5), в то время как градиентная поправка, которая более аккуратно зануляет параметры, не создает провал логарифма правдоподобия.

4.5.2 Значения R

Регуляризатор R и коэффициент τ вводились из соображений, что требуется найти такое решение задачи, в котором $\tau R \to \max$. Поскольку в рассматриваемом случае $\tau < 0$, то требуется $R \to \min$.

Рис. 4: $|T| = 3$. Значения R очень малы по сравенению с L, поэтому при малых значениях $|\tau|$ оказывается, что оптимизировать L существенно выгоднее, чем R, и можно наблюдать немонотонную зависимость R от итераций. Посколько значения R практически не влияют на оптимизацию, то эффект скачков выражен неярко. Градиентные поправки заметно хуже оптимизируют R.

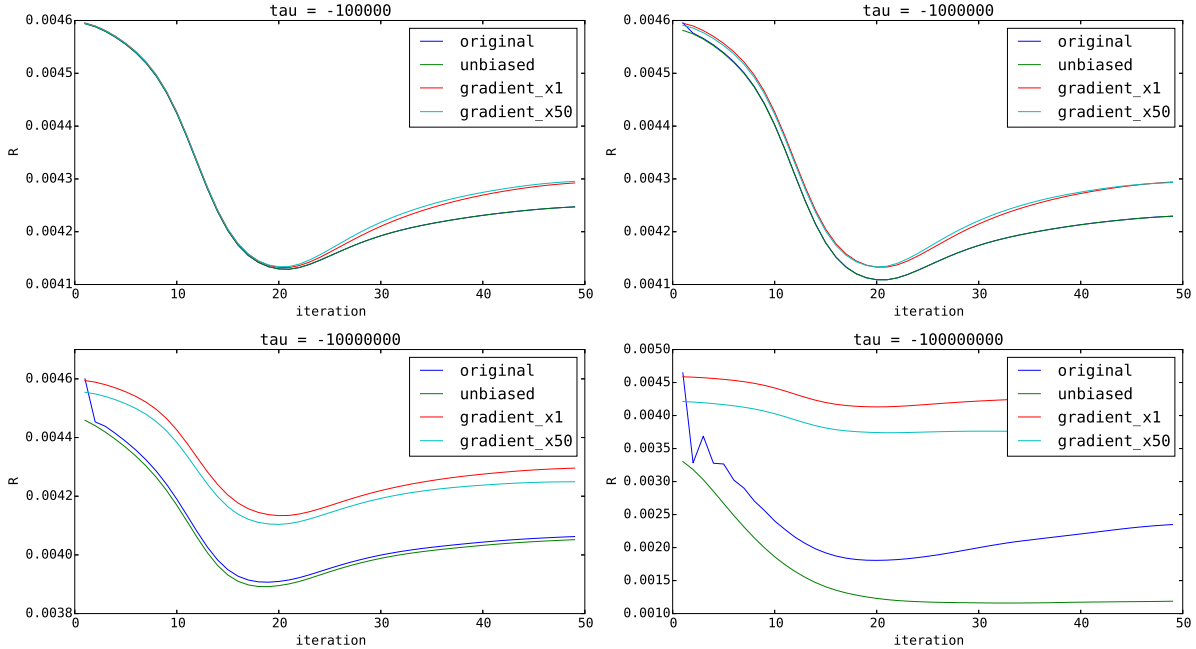


Рис. 5: $|T| = 10$. На верхних графиках виден скачок R на первой итераций, на нижних графиках он ярче выражен. Несмещённая модификация и градиент ×50 лучше всех остальных.

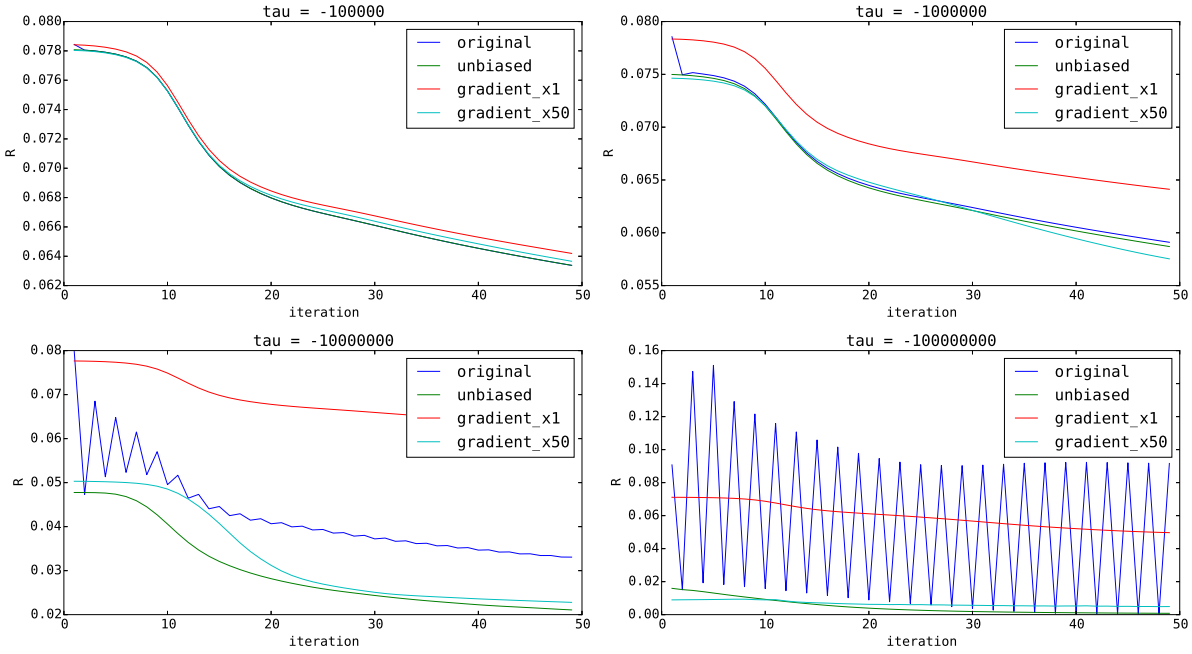
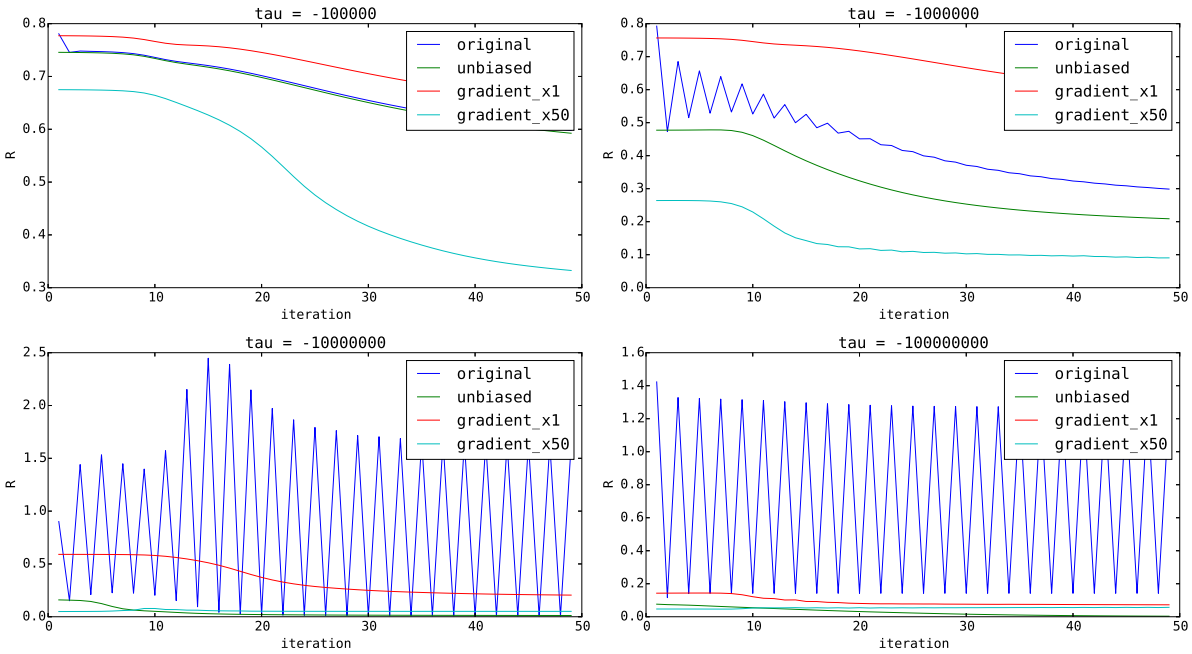


Рис. 6: $|T| = 30$. Эффект скачков уже ярко выражен при $\tau = -10^6$. Градиент ×50 существенно лучше всех остальных.



При достаточно больших значения $|\tau|$ видно, что в стандартной формуле М-шага есть скачки в изменении R на первых итерациях. Это вызвано тем, что регуляризационные поправки считаются в точке с предыдущей итерации. На первых итерациях она слабо связана с точкой $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$, к которой применяется регуляризационное пре-

образование. Конечно, со временем $(\varphi_{wt}, \theta_{td})$ стремятся к $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$. Это можно видеть на графиках — они постепенно спрямляются. Но на первых итерациях $(\varphi_{wt}, \theta_{td})$ существенно отличается от $\left(\frac{n_{wt}}{n_t}, \frac{n_{td}}{n_d}\right)$ и эффект скачков проявляется, а большое значение $|\tau|$ усиливает различие и делает эффект более ярким.

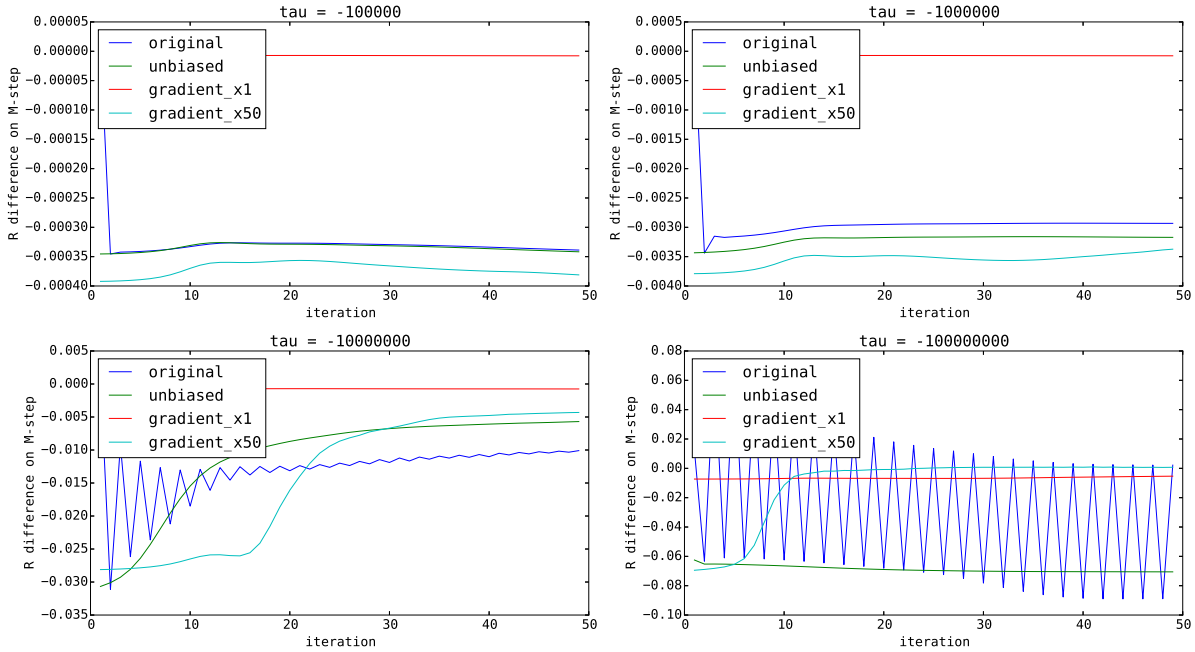
С ростом числа тем порядок значений R изменился, и поэтому чувствительность к $|\tau|$ возросла. Как следствие, можно наблюдать, что колебания значений функционала начинаются раньше, размер этих колебаний существеннее, а градиентные поправки раньше начинают показывать лучшее качество, чем остальные алгоритмы.

Также стоит отметить интересный эффект при $|T| = 3$. Значения R очень малы, и в итоге оптимизация L выходит на первый план, что приводит к немонотонной траектории R. С ростом τ эта зависимость постепенно выпрямляется. Это хорошо иллюстрирует, как коэффициент регуляризации влияет на модель, увеличивая важность регуляризатора.

4.5.3 Изменения R на втором этапе M-шага

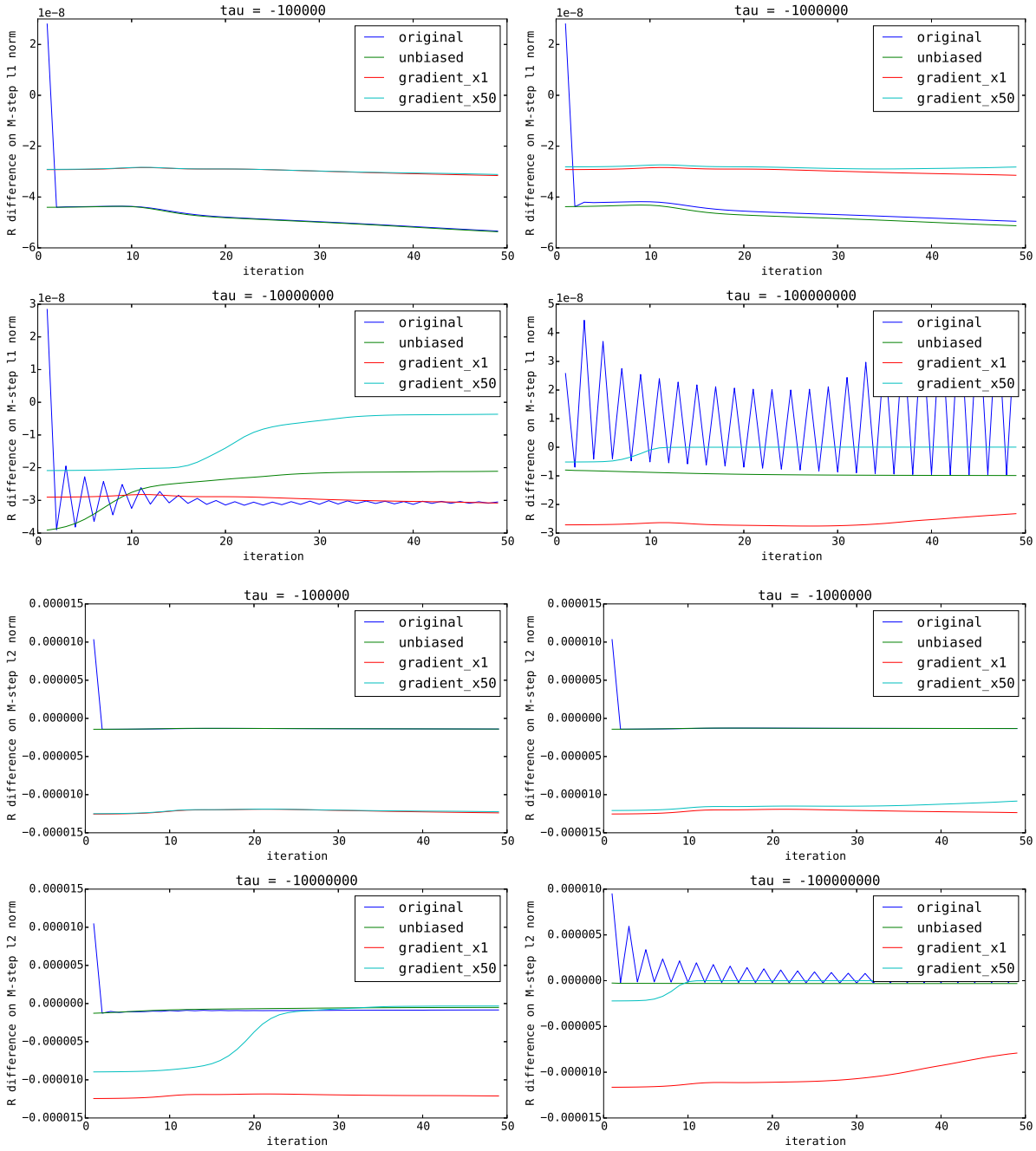
Предложенные формулы выводились из соображений увеличения функционала R при регуляризационном преобразовании. Чтобы проверить правильность выведенных формул, предлагается измерить изменение R на данном этапе M-шага. Поскольку для выбранного регуляризатора стоит задача минимизации R, то чем меньше изменение, тем лучше. Также стоит отметить, что эффект скачков на первых итерациях будет здесь наиболее заметен, так как проявляется именно в данной величине, а все предыдущие замеры совмещали данный эффект с другими.

Рис. 7: $|T| = 10$. Для несмещённой и стандартной формулы изменение существенно лучше, что подтверждает слова о неправильном выборе констант перед градиентом.



Другие графики аналогичны, поэтому не будут приведены. Чтобы сравнить формулы по изменению R при регуляризационном преобразовании, предложено оценить удельный эффект (нормировать изменение на l_1 и l_2 нормы). Графики весьма однотипны, поэтому приведём только случай $|T| = 10$.

Рис. 8: $|T| = 10$. Стандартная формула и несмещённая модификация лучше по l_1 норме, градиентные поправки по l_2 норме. Эффект скачков хорошо выражен.



Для всех предложенных модификаций уменьшение значения R на втором этапе М-шага заметно больше, чем в стандартном алгоритме, что ожидаемо, поскольку они были выведены с такой целью. Поскольку градиент — это оптимальное направление изменения в l_2 норме, то при нормировке изменения R на l_2 норму получаем, что градиентные поправки существенно эффективнее двух других методов.

С ростом $|\tau|$ можно наблюдать эффект насыщения, то есть эффективность поправок падает с итерациями, что означает приближение к стационарной точке R, это положительно говорит о предложенных модификациях.

4.5.4 Минимальное значения в Φ и Θ

Важным свойством при доказательстве сходимости была ε-разреживаемость регуляризатора, поэтому требуется проверить данное предположение. На графиках изображены логарифмы минимальных ненулевых значений в матрицах Φ и Θ.

Рис. 9: $|T| = 3$. Для матрицы Φ градиентные поправки имеют большую отделимость от нуля. На матрице Θ различия несущественны при отсутсвии эффекта скачков и сверхбольшого $|\tau|$.

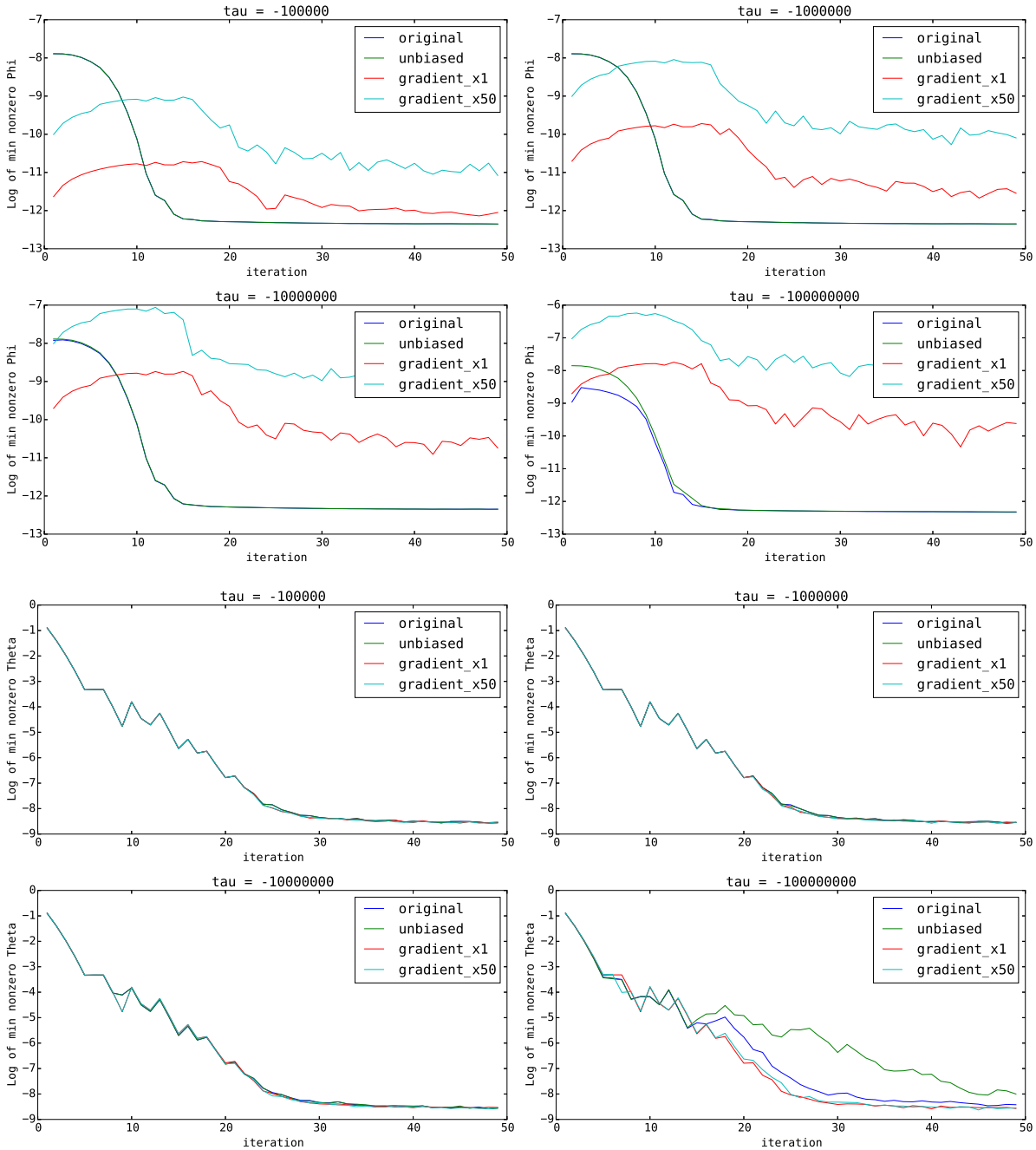


Рис. 10: $|T| = 10$. Графики несмещённой модификации и стандартной формулы стабилизируются на некотором значении, это вызвано технической особенностью, описанной в главе 4.4.

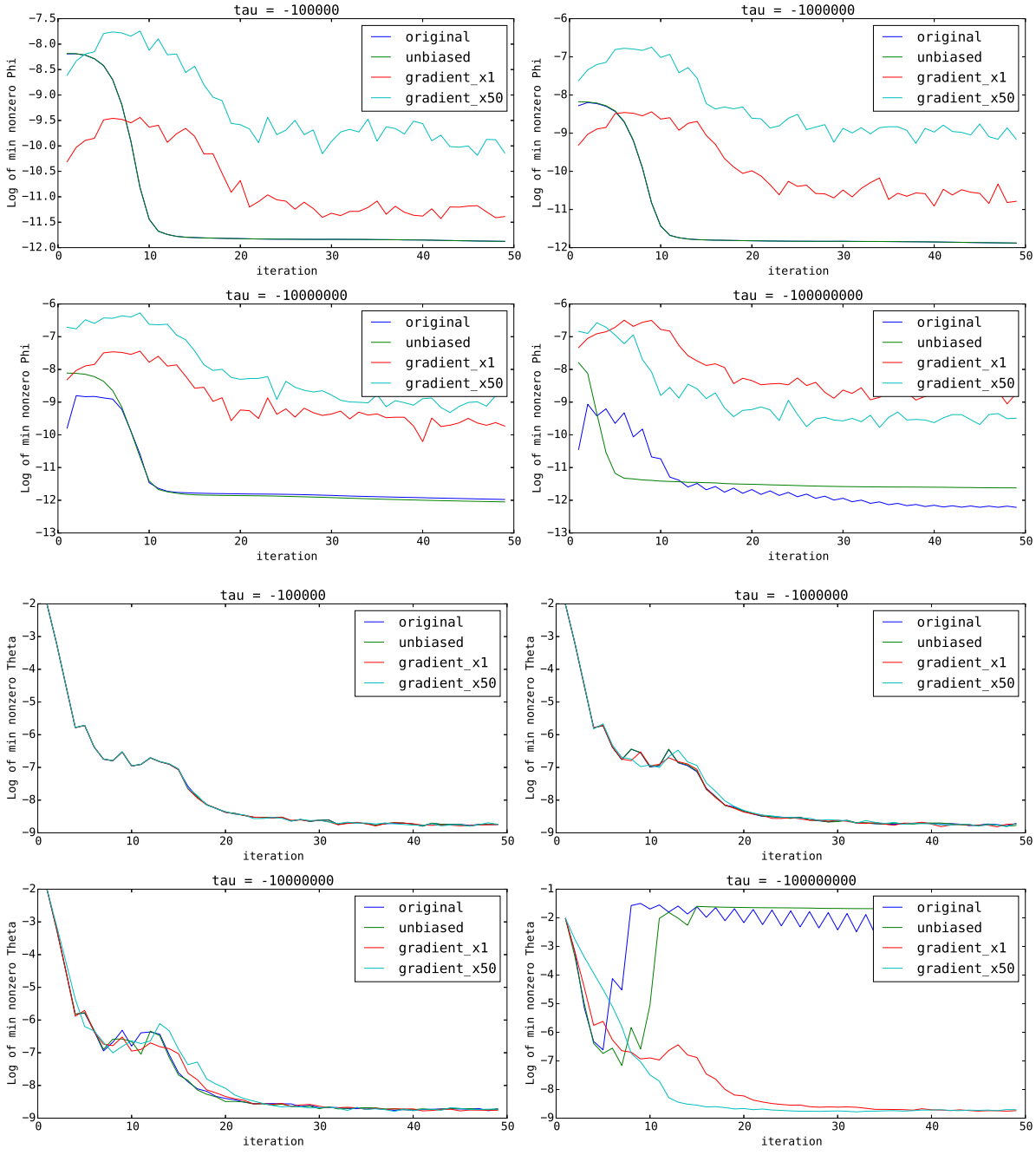
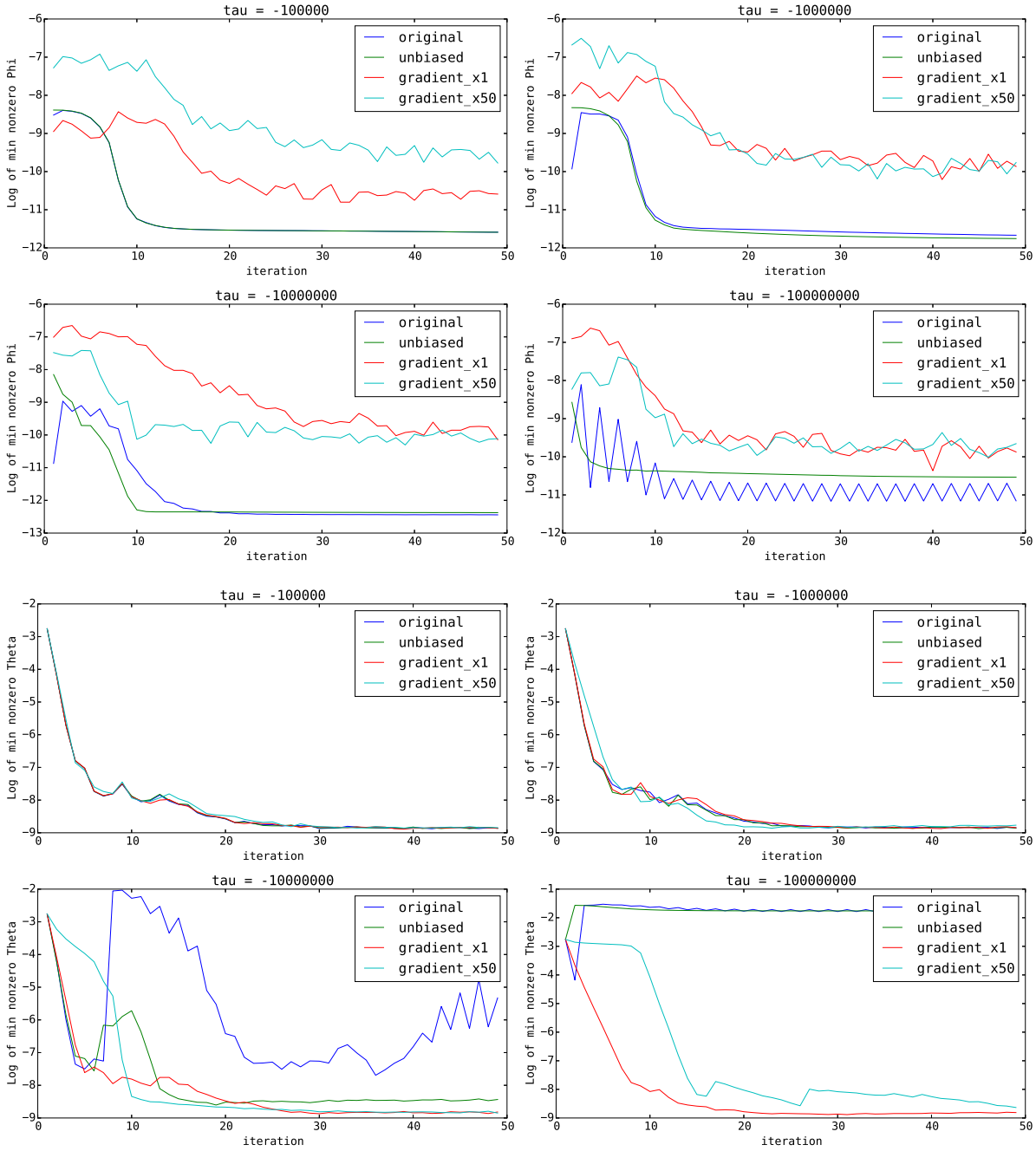


Рис. 11: $|T| = 30$. Так же как и в предыдущих в графиках, дестабилизация происходит раньше из-за увеличения порядка значений R.



Несмотря на то, что градиентные методы аккуратнее зануляют элементы Φ, они делают это заметно лучше – значения существенно сильнее отделимы от нуля. Однако, в элементах Θ разницы нет, если не считать случая очень больших $|\tau|$, когда сильные множественные зануления привели к существенной отделимости параметров от нуля.

4.5.5 Минимальный размер темы

Второе важное свойство для сходимости алгоритма — это сильная регулярность регуляризатора 3.2. Фактически, важна отделимость от нуля знаменателя при нормировке в M-шаге. Именно эти значения и были получены.

Рис. 12: $|T| = 3$.

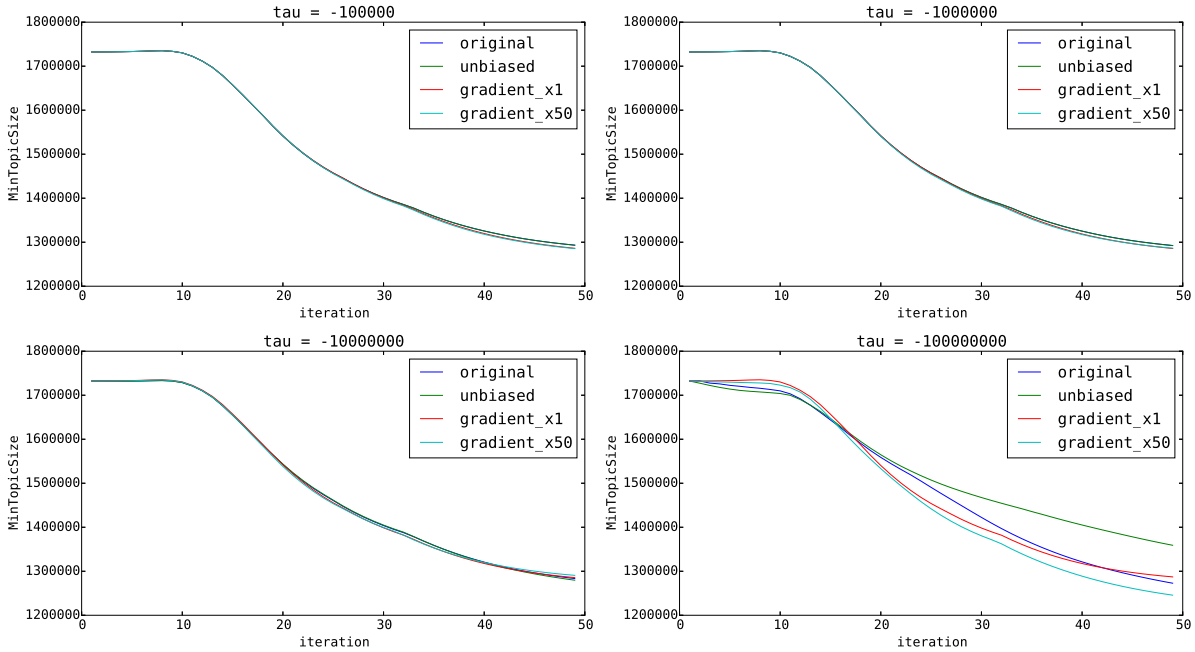


Рис. 13: $|T| = 10$.

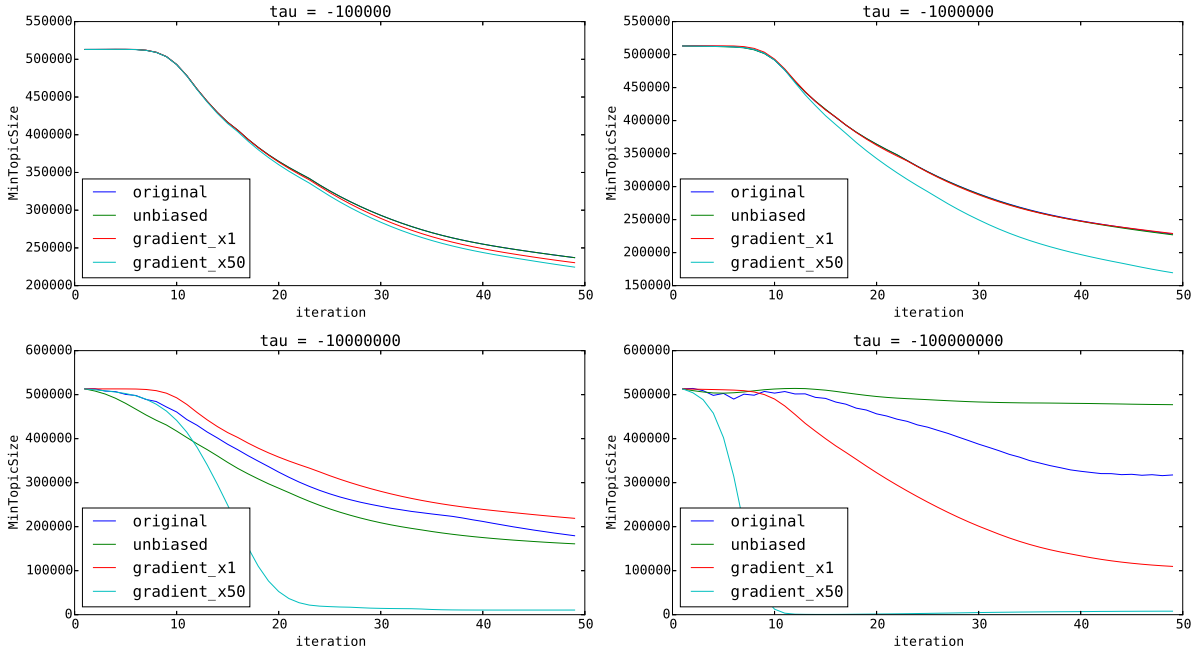
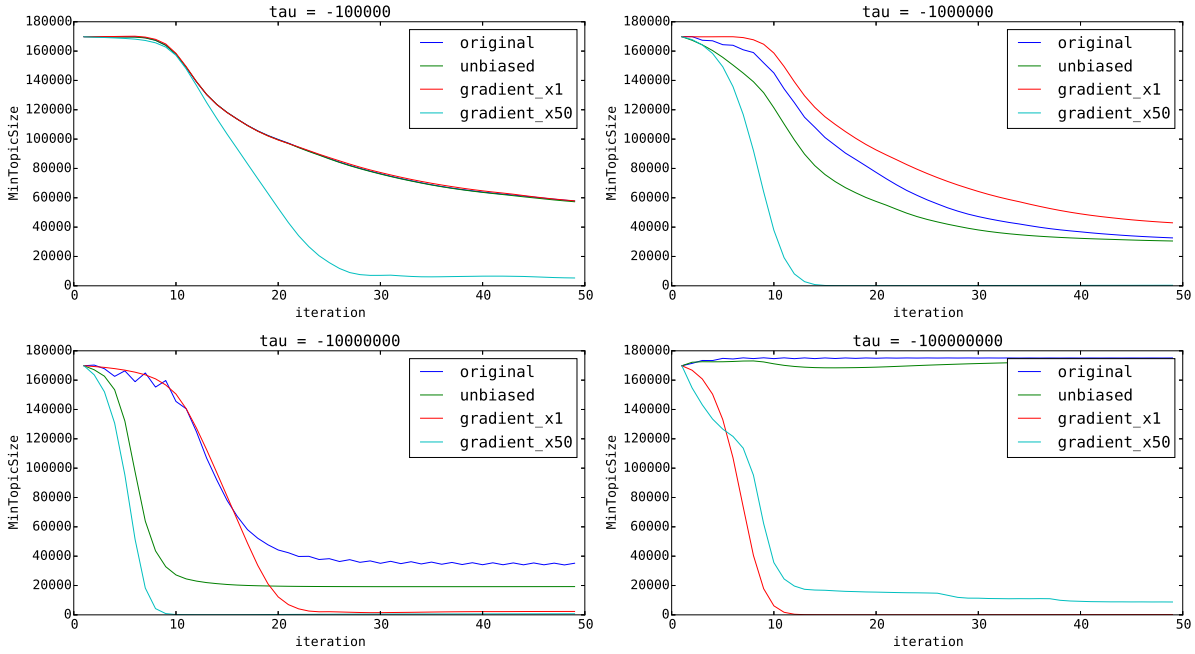


Рис. 14: $|T| = 30$.



Значения становятся очень малыми, но отделимость от нуля сохраняется (просто отделено небольшой константой порядка 1, но на фоне 10000 это кажется нулём).

Также видно, что градиентные методы более склонны к селекции тем. Отсюда можно сделать вывод, что градиентное направление обладает большей селективностью, чем стандартная формула. Также стоит отметить, что несмещённая модификация активнее отбирает темы чем стандартная формула.

4.5.6 Итоги экспериментов

На основании проведённых экспериментов можно сделать следующие выводы:

1. Предположения о ε-отделимости и сильной регулярности выполняются на практике, или могут легко гарантироваться за счёт реализации.

2. С точки зрения оптимизации $L + \tau R$ все рассмотренные формулы отличаются незначительно. Основное достоинтсво предложенных модификаций в этом плане — это теоретические гарантии.

3. Есть эффект скачков значений функционалов на первых итерациях для стандартной формулы. Его можно избежать, если пользоваться несмещённой модификацией, для которой есть доказанное утверждение про увеличение R, что приводит к более гладкой траектории. Таким образом, данная модификация явно улучшает стандартную формулу.

4. Для градиентных поправок необходимо дополнительное исследование, чтобы понять как подбирать константу перед градиентом. Эксперименты показали наличие потенциала подобной модификации.

5 Заключение

Подведём краткое резюме данной работы. В вероятностном тематическом моделированиии существует подход ARTM, он предоставляет быстрый и очень гибкий функционал для оптимизации, легко адаптируемый под конкретную задачу. Основная его идея состоит в максимизации регуляризованного правдоподобия при помощи EM-алгоритма. На практике предложенный алгоритм успешно сходился, однако, не было теоретического обоснования данной сходимости. Определение ограничений, при выполнении которых, можно гарантировать сходимость, было целью данной работы. Итерации алгоритма ARTM были проинтерпретированы как итерации GEM алгоритма, для которых условия сходимости хорошо изучены. Используя данные результаты, были найдены естественные ограничения на регуляризатор, выполнение которых достаточно просто проверить на практике. Пусть k — это номер итерации, тогда полученные условия можно сформулировать следующим образом:

1. Сохранение нуля регуляризатором.

   $n_{wt}^k = 0 \Rightarrow \varphi_{wt}^k = 0$, $n_{td}^k = 0 \Rightarrow \theta_{td}^k = 0$.

2. ε-отделимость от нуля элементов матриц Φ и Θ.

   $\exists \varepsilon > 0 \ \exists N \ \forall k > N \ \varphi_{wt}^k, \theta_{td}^k \notin (0, \varepsilon)$.

3. Конечность логарифма правдоподобия на итерациях.

   $n_{dw} > 0 \Rightarrow \forall k \ \exists t: p_{tdw}^k > 0$.

4. Отделимость от нуля знаменателя на M-шаге.

   $\exists \delta > 0 \ \exists N \ \forall k > N \ \forall t \ \exists w \ n_{wt}^k + r_{wt}^k > \delta$ и аналогичное условие для θ.

5. Неуменьшение нижней оценки регуляризованного правдоподобия на итерациях.

   $\exists N \ \forall k > N$: $Q^k(\varphi^k, \theta^k) + R(\Phi^k, \Theta^k) \geqslant Q^k(\varphi^{k-1}, \theta^{k-1}) + R(\Phi^{k-1}, \Theta^{k-1})$, где $Q^k(\varphi, \Theta) = \sum\limits_{t,d,w} p_{tdw}^k (\ln \varphi_{wt} + \ln \theta_{td})$.

Первое условие легко проверяется аналитически. Второе и третье условия можно обеспечить при реализации алгоритма. Четвёртое условие для матрицы Φ можно проинтерпретировать как критерий селекции тем. То есть, если значение становится меньше δ, то зануляется вся строчка матрицы Φ. С точки зрения реализации это эквивалентно просто удалению строчки в матрице и уменьшению числа тем. Для матрицы Θ выполнение данного условия можно достичь за счёт выбора τ. Пятое условие должно быть обеспеченно за счёт правильного выбора регуляризационных добавок. Стандартный алгоритм предлагает использовать $r_{wt}^k = \tau\varphi_{wt}^{k-1}\frac{\partial R}{\partial\varphi_{wt}}(\varphi_{wt^{k-1}}, \theta_{td}^{k-1})$ и $r_{td}^k = \tau\theta_{td}^{k-1}\frac{\partial R}{\partial\theta_{td}}(\varphi_{wt}^{k-1}, \theta_{td}^{k-1})$. Для данных формул не удалось получить хороших оценок, поэтому была рассмотрена следующая модификация: замена всех вхождений φ_{wt} и θ_{td} на их несмещённые оценки. То есть $r_{wt}^k = \tau\frac{n_{wt}^k}{n_t^k}\frac{\partial R}{\partial\varphi_{wt}}\big(\frac{n_{wt}^k}{n_t^k}, \frac{n_{td}^k}{n_d^k}\big)$ и $r_{td}^k = \tau\frac{n_{td}^k}{n_d^k}\frac{\partial R}{\partial\theta_{td}}\big(\frac{n_{wt}^k}{n_t^k}, \frac{n_{td}^k}{n_d^k}\big)$. Используя идею, что можно рассматривать функцию R не как функцию от φ_{wt} и θ_{td}, а как функцию от n_{wt} и n_{td}, только с нормировкой аргументов, было получено, что на М-шаге происходит увеличение R, если взято не слишком большое τ. Также была предложена идея использовать вычисленный градиент R в качестве регуляризационных добавок на М-шаге: $r_{wt}^k = \tau A_t\big[\frac{\partial R}{\partial\varphi_{wt}} - \sum\limits_u \varphi_{ut}\frac{\partial R}{\partial\varphi_{ut}}\big]\big(\frac{n_{wt}^k}{n_t^k}, \frac{n_{td}^k}{n_d^k}\big)$ и $r_{td}^k = \tau B_d\big[\frac{\partial R}{\partial\theta_{td}} - \sum\limits_s \theta_{sd}\frac{\partial R}{\partial\theta_{sd}}\big]\big(\frac{n_{wt}^k}{n_t^k}, \frac{n_{td}^k}{n_d^k}\big)$. Определение формул для A_t и B_d в данной работе не производилось, для экспериментов использовались самые наивные варианты.

Был проведён эксперимент, в котором сравнивались три возможных формулы М-шага. Предложенные модификации показали небольшое улучшение по сравнению со стандартными формулами. Также был замечен следующий эффект: при достаточно больших τ наблюдаются скачки в функционалах для стандартной формулы М-шага, это свидетельствует в пользу того, что не получится теоретически доказать неуменьшение R на М-шаге для стандартной формулы.

5.1 Результаты, выносимые на защиту

1. Условия для сходимости EM-алгоритма ARTM, легко проверяемые и обеспечиваемые при реализациии.

2. Две модификации формул М-шага EM-алгоритма, улучшающие сходимость без увеличения вычислительной сложности.

3. Оценки изменения значений регуляризатора и логарифма правдоподобия на итерациях для предложенных модификаций.

Список литературы

[1] *Andrzejewski D., Buttler D.* Latent topic feedback for information retrieval // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining / ACM. — 2011. — Pp. 600–608.

[2] *Bilmes J. A. et al.* A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models // *International Computer Science Institute.* — 1998. — Vol. 4, no. 510. — P. 126.

[3] *Blei D. M., Lafferty J. D.* A correlated topic model of science // *The Annals of Applied Statistics.* — 2007. — Pp. 17–35.

[4] *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation // *the Journal of machine Learning research.* — 2003. — Vol. 3. — Pp. 993–1022.

[5] *Bolelli L., Ertekin Ş., Giles C. L.* Topic and trend detection in text collections using latent dirichlet allocation // Advances in Information Retrieval. — Springer, 2009. — Pp. 776–780.

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

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

[8] Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora / J. Zhang, Y. Song, C. Zhang, S. Liu // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining / ACM. — 2010. — Pp. 1079–1088.

[9] *Gilks W. R., Richardson S., Spiegelhalter D. J.* Introducing markov chain monte carlo // *Markov chain Monte Carlo in practice.* — 1996. — Vol. 1. — P. 19.

[10] *Griffiths T. L., Steyvers M.* Finding scientific topics // *Proceedings of the National Academy of Sciences.* — 2004. — Vol. 101, no. suppl 1. — Pp. 5228–5235.

[11] *Gruber A., Weiss Y., Rosen-Zvi M.* Hidden topic markov models // International conference on artificial intelligence and statistics. — 2007. — Pp. 163–170.

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

[13] An introduction to mcmc for machine learning / C. Andrieu, N. De Freitas, A. Doucet, M. I. Jordan // *Machine learning.* — 2003. — Vol. 50, no. 1-2. — Pp. 5–43.

[14] An introduction to variational methods for graphical models / M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul // *Machine learning.* — 1999. — Vol. 37, no. 2. — Pp. 183–233.

[15] Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou, F. Muhammad // *Frontiers of computer science in China.* — 2010. — Vol. 4, no. 2. — Pp. 280–301.

[16] Lcars: A spatial item recommender system / H. Yin, B. Cui, Y. Sun et al. // *ACM Transactions on Information Systems (TOIS).* — 2014. — Vol. 32, no. 3. — P. 11.

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

[18] Modeling location-based user rating profiles for personalized recommendation / H. Yin, B. Cui, L. Chen et al. // *ACM Transactions on Knowledge Discovery from Data (TKDD).* — 2015. — Vol. 9, no. 3. — P. 19.

[19] *Nallapati R., Cohen W. W.* Link-plsa-lda: A new unsupervised model for topics and influence of blogs // ICWSM. — 2008.

[20] Originator or propagator?: incorporating social role theory into topic models for twitter content analysis / X. W. Zhao, J. Wang, Y. He et al. // Proceedings of the 22nd ACM international conference on Conference on information & knowledge management / ACM. — 2013. — Pp. 1649–1654.

[21] Probabilistic author-topic models for information discovery / M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining / ACM. — 2004. — Pp. 306–315.

[22] Statistical topic models for multi-label document classification / T. N. Rubin, A. Chambers, P. Smyth, M. Steyvers // *Machine learning.* — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.

[23] Textflow: Towards better understanding of evolving topics in text / W. Cui, S. Liu, L. Tan et al. // *Visualization and Computer Graphics, IEEE Transactions on.* — 2011. — Vol. 17, no. 12. — Pp. 2412–2421.

[24] *Topsøe F.* Some inequalities for information divergence and related measures of discrimination // *Information Theory, IEEE Transactions on.* — 2000. — Vol. 46, no. 4. — Pp. 1602–1609.

[25] *Varshney D., Kumar S., Gupta V.* Modeling information diffusion in social networks using latent topic information // Intelligent Computing Theory. — Springer, 2014. — Pp. 137–148.

[26] *Vorontsov K.* Additive regularization for topic models of text collections // Doklady Mathematics / Citeseer. — Vol. 89. — 2014. — Pp. 301–304.

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

[28] *Vorontsov K., Potapenko A.* Additive regularization of topic models // *Machine Learning.* — 2015. — Vol. 101, no. 1-3. — Pp. 303–323.

[29] *Vulić I., De Smet W., Moens M.-F.* Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora // *Information Retrieval.* — 2013. — Vol. 16, no. 3. — Pp. 331–368.

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

[31] *Wallach H. M., Mimno D. M., McCallum A.* Rethinking lda: Why priors matter // Advances in neural information processing systems. — 2009. — Pp. 1973–1981.

[32] *Wang H., Zhang D., Zhai C.* Structural topic model for latent topical structure analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 / Association for Computational Linguistics. — 2011. — Pp. 1526–1535.

[33] *Wu C. J.* On the convergence properties of the em algorithm // *The Annals of statistics.* — 1983. — Pp. 95–103.

[34] *Zhou S., Li K., Liu Y.* Text categorization based on topic model // *International Journal of Computational Intelligence Systems.* — 2009. — Vol. 2, no. 4. — Pp. 398–409.