Регуляризация, робастность и разреженность вероятностных тематических моделей

Константин Воронцов
Вычислительный Центр им. А. А. Дородницына РАН

Анна Потапенко
Факультет ВМК МГУ им. М. В. Ломоносова

Интеллектуализация Обработки Информации, ИОИ-9
16–22 сентября 2012, Черногория, г. Будва

Содержание

1. **Задача тематического моделирования**
   - Постановка задачи и вероятностные предположения
   - Модель PLSA и EM-алгоритм
   - Эмпирические оценки качества тематических моделей
2. **От модели PLSA к модели LDA**
   - Эвристика частого обновления параметров
   - Эвристика сэмплирования
   - Эвристика регуляризации
3. **Робастная вероятностная тематическая модель**
   - Модель с фоновой и шумовой компонентами
   - EM-алгоритм для робастной модели
   - Разреженность робастной модели

Задача определения тематики коллекции документов

Дано:

W — словарь, множество слов (терминов);

D — множество (коллекция, корпус) текстовых документов;

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$.

Найти:

— к каким темам относится каждый документ

— какими терминами определяется каждая тема

— сколько тем содержится в коллекции

Гипотеза «мешка слов»:

порядок терминов не важен для определения тематики текста.

Гипотеза разреженности:

документ, как правило, относится к небольшому числу тем;

тема, как правило, определяется небольшим числом терминов.

Вероятностная формализация постановки задачи

Вероятностные предположения:

- каждое слово в документе связано с некоторой темой $t \in T$;
- коллекция D — это выборка независимых наблюдений (d, w) из дискретного распределения $p(d, w, t)$ на $D \times W \times T$;
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$;

Вероятностная модель порождения документа d:

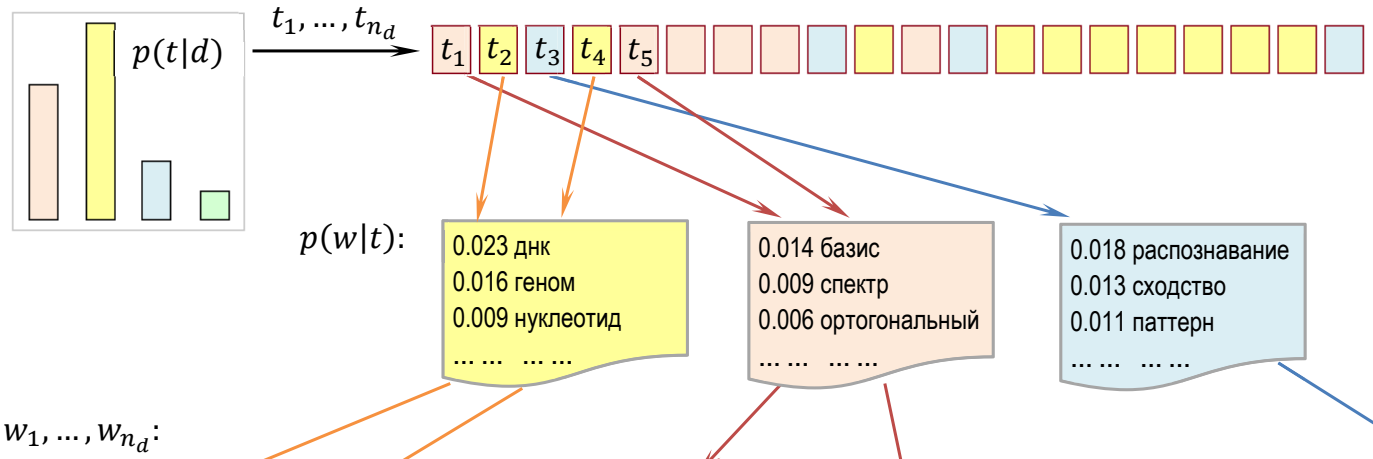
$$p(w|d) = \sum_{t \in T} p(w|t)\, p(t|d)$$

Найти:

- $p(w|t)$ — распределение терминов в каждой теме $t \in T$;
- $p(t|d)$ — распределение тем в каждом документе $d \in D$.

Вероятностная тематическая модель порождения документа d

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Частотные оценки условных вероятностей

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$$

Если рассматривать коллекцию как выборку троек (d, w, t), то

$$\hat{p}(w|d) = \frac{n_{dw}}{n_d}, \qquad \hat{p}(w|t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t|d) = \frac{n_{dt}}{n_d};$$

n_{dwt} — число троек (d, w, t) во всей коллекции;

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

$n_{dt} = \sum_{w \in d} n_{dwt}$; $\qquad n_d = \sum_{w \in d} \sum_{t \in T} n_{dwt}$ — длина документа d;

$n_{wt} = \sum_{d \in D} n_{dwt}$; $\qquad n_t = \sum_{d \in D} \sum_{w \in d} n_{dwt}$ — «длина темы» t;

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

Цели тематического моделирования (topic modeling)

- Тематический поиск документов по тексту любой длины
- Категоризация, классификация, аннотирование, суммаризация текстовых документов
- Тематический поиск объектов, связанных с документами: рисунков, авторов, организаций, журналов, конференций
- Выявление трендов и фронта исследований

Типичные приложения:

- Поиск научной информации
- Анализ и агрегирование новостных потоков
- Рекомендательные сервисы (коллаборативная фильтрация)
- Рубрикация коллекций изображений, видео, музыки
- Аннотация генома и другие задачи биоинформатики

Вероятностный латентно-семантический анализ
PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Максимизация правдоподобия по $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$:

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

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{d \in D} \theta_{td} = 1.$$

Это задача неотрицательного матричного разложения $F \approx \Phi\Theta$,

$F = \big(\hat{p}(w|d)\big)_{W \times D}$ — известная матрица исходных данных;
$\Phi = \big(\phi_{wt}\big)_{W \times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$;
$\Theta = \big(\theta_{td}\big)_{T \times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

EM-алгоритм

E-шаг: если ϕ_{wt}, θ_{td} известны, то по формуле Байеса вычисляются условные вероятности тем $t \in T$ для всех (d, w):

$$H_{dwt} \equiv p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}. \qquad (1)$$

M-шаг: если H_{dwt} известны, то решение задачи максимизации правдоподобия аналитически выражается через H_{dwt}:
(по сути, это частотные оценки условных вероятностей)

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_{wt} = \sum_{d \in D} n_{dw} H_{dwt}, \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt};$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_{dt} = \sum_{w \in d} n_{dw} H_{dwt}, \qquad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}.$$

EM-алгоритм — это чередование E и M шагов до сходимости.

Рационализация EM-алгоритма: E-шаг встроен внутрь M-шага

Идея: не хранить H_{dwt}, а вычислять по мере необходимости.
Сложность алгоритма $O(|D| \cdot |W| \cdot |T|)$.

Вход: коллекция D, число тем $|T|$, начальные Φ и Θ;
Выход: распределения Φ и Θ;

повторять
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D$, $w \in W$, $t \in T$;
  для всех $d \in D$, $w \in d$
    $Z := \sum_t \phi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\phi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $n_{dw} \cdot \frac{1}{Z}\phi_{wt}\theta_{td}$;
  $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D$, $t \in T$;
пока Φ и Θ не стабилизируются.

Стандартная методика оценивания тематических моделей

Перплексия тестовой коллекции D' (hold-out perplexity):

$$\mathcal{P}(D') = \exp\left(-\frac{1}{n}\sum_{d\in D'}\sum_{w\in d''} n_{dw}\ln p(w|d)\right),$$

n — длина коллекции;
$d = d' \sqcup d''$ — случайное разбиение контрольного документа на две половины равной длины;

параметры ϕ_{wt} оцениваются по обучающей коллекции D;
параметры θ_{td} оцениваются по первой половине d';
перплексия вычисляется по второй половине d''.

Интерпретации перплексии:
1) $\mathcal{P}(D') \to |W|$ при $n \to \infty$, если слова равновероятны;
2) насколько хорошо мы можем предсказывать появление слов (чем меньше перплексия, тем лучше).

Другие методики оценивания тематических моделей

- Число ошибок классификации размеченной тестовой коллекции D'.
- Отклонение от *гипотезы условной независимости* $p(w|d,t) = p(w|t)$ на обучающей коллекции D для темы t:

$$\mathrm{KL}\Big(\hat{p}(d,w|t),\ \hat{p}(d|t)\cdot\hat{p}(w|t)\Big) = \sum_{d,w}\frac{n_{dwt}}{n_t}\log\frac{n_{dwt}\cdot n_t}{n_{td}\cdot n_{wt}}$$

D.Mimno, D.Blei. Bayesian Checking for Topic Models // Empirical Methods in Natural Language Processing, 2011.

- Доля случаев, когда эксперт верно определяет:
  — лишнюю тему в списке главных тем документа;
  — лишний термин в списке главных терминов темы.

J.Chang, J.Boyd-Graber, S.Gerrish, C.Wang, D.Blei. Reading Tea Leaves: How Humans Interpret Topic Models // Advances in Neural Information Processing Systems 22, 2009, pp. 288–296.

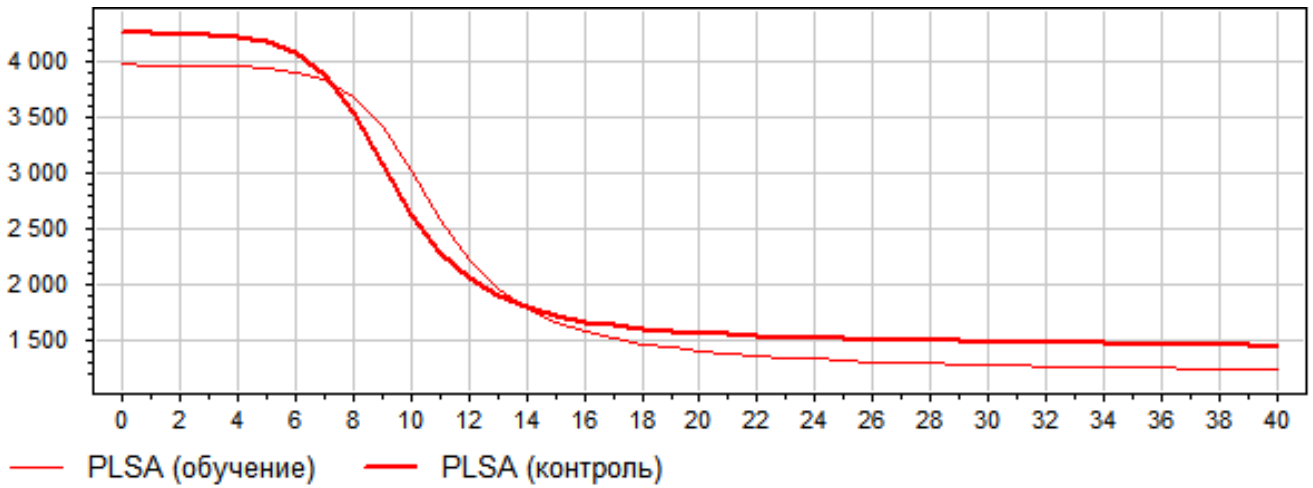
Методика эксперимента

D — коллекция 2000 авторефератов диссертаций на русском языке суммарной длины $n \approx 8.7 \cdot 10^6$, словарь $|W| \approx 3 \cdot 10^4$.

Предобработка: лемматизация, удаление стоп-слов.

D' — коллекция 200 авторефератов, не включённых в D.

Строятся графики зависимости перплексии от числа итераций (проходов коллекции); число итераций 40; число тем $|T| = 100$;



Недостатки PLSA

- PLSA переобучается, т.к. параметров ϕ_{wt} и θ_{td} слишком много ($|D| \cdot |T| + |W| \cdot |T|$), и на них не накладывается никаких ограничений регуляризации.
- PLSA неверно оценивает вероятность новых слов:
  если $n_w = 0$, то $\hat{p}(w|t) = 0$ для всех $t \in T$.
- PLSA не позволяет управлять *разреженностью* Φ и Θ, т.к.
  (в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в финале $\phi_{wt} = 0$);
  (в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в финале $\theta_{td} = 0$)

Гипотеза разреженности матриц Φ и Θ:
каждый документ относится к небольшому числу тем;
каждая тема описывается небольшим числом терминов...

т.е. на самом деле параметров должно быть намного меньше.

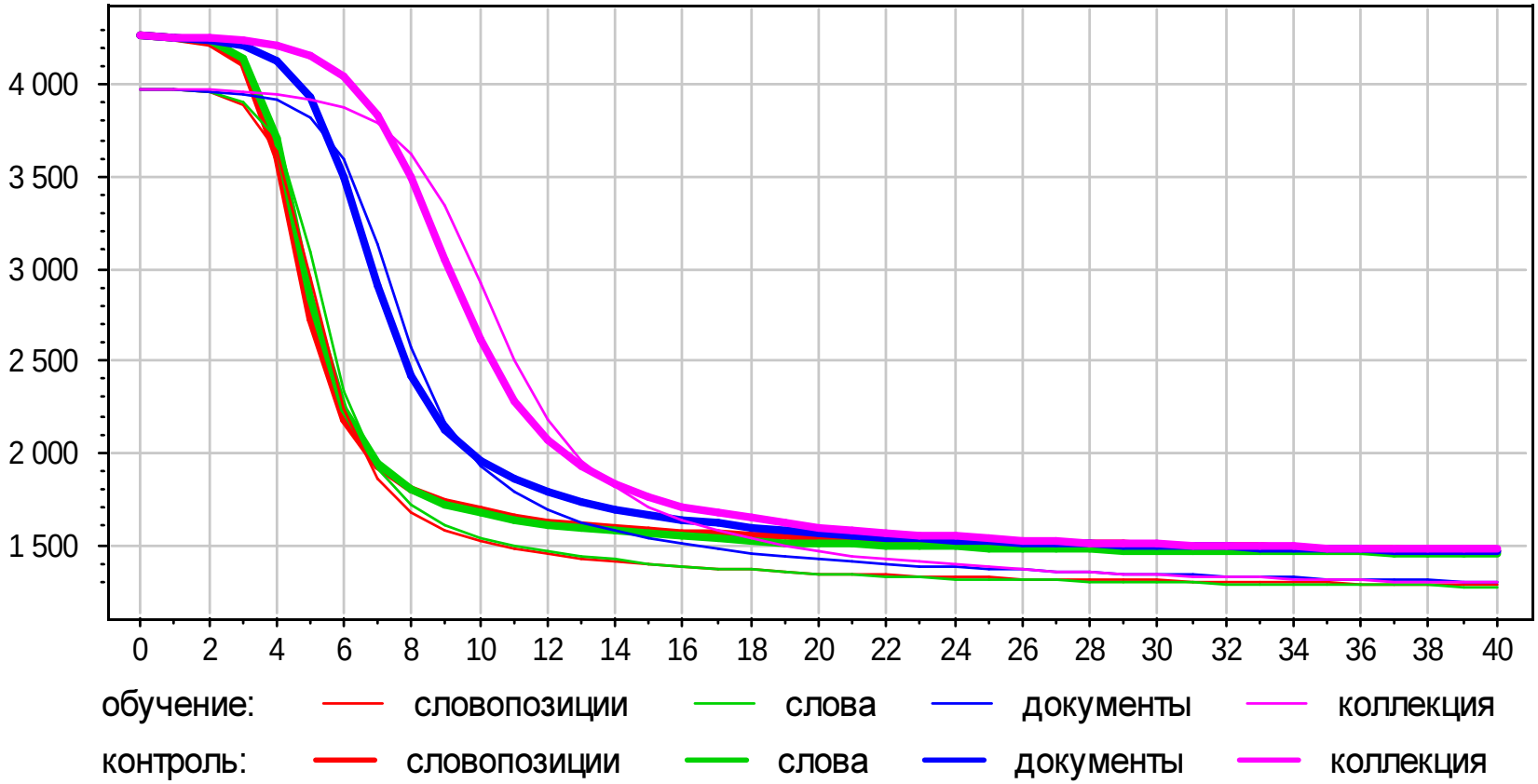
Обобщённый EM-алгоритм (GEM, generalized EM-algorithm)

Идея: не обязательно точно решать задачу M-шага, достаточно сместиться в направлении максимума.
В PLSA это приводит к *частым обновлениям параметров* Φ, Θ:

Вход: коллекция D, число тем $|T|$, начальные Φ и Θ;
Выход: распределения Φ и Θ;

обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$, n_{dwt} для всех $d \in D$, $w \in W$, $t \in T$;
повторять
  для всех $d \in D$, $w \in d$
    для всех $t \in T$ таких, что $n_{dwt} > 0$ или $\phi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\hat{n}_d$ на $(n_{dw}H_{dwt} - n_{dwt})$;
      $n_{dwt} := n_{dw}H_{dwt}$;
    если не первая итерация и пора обновить Φ, Θ **то**
      $\phi_{wt} := \hat{n}_{wt}/\hat{n}_t$; $\theta_{td} := \hat{n}_{dt}/\hat{n}_d$;
пока Φ и Θ не стабилизируются.

Частота обновления Φ и Θ не влияет на качество модели



Частота обновления параметров Φ и Θ не влияет на качество, а только на скорость сходимости.

Вывод: лучше обновлять после каждого слова (d, w).

Проблема сокращения объёма промежуточных данных

В PLSA число n_{dw} пропорционально распределяется между темами, и приходится хранить числа $n_{dwt} = n_{dw} H_{dwt}$, $t \in T$.

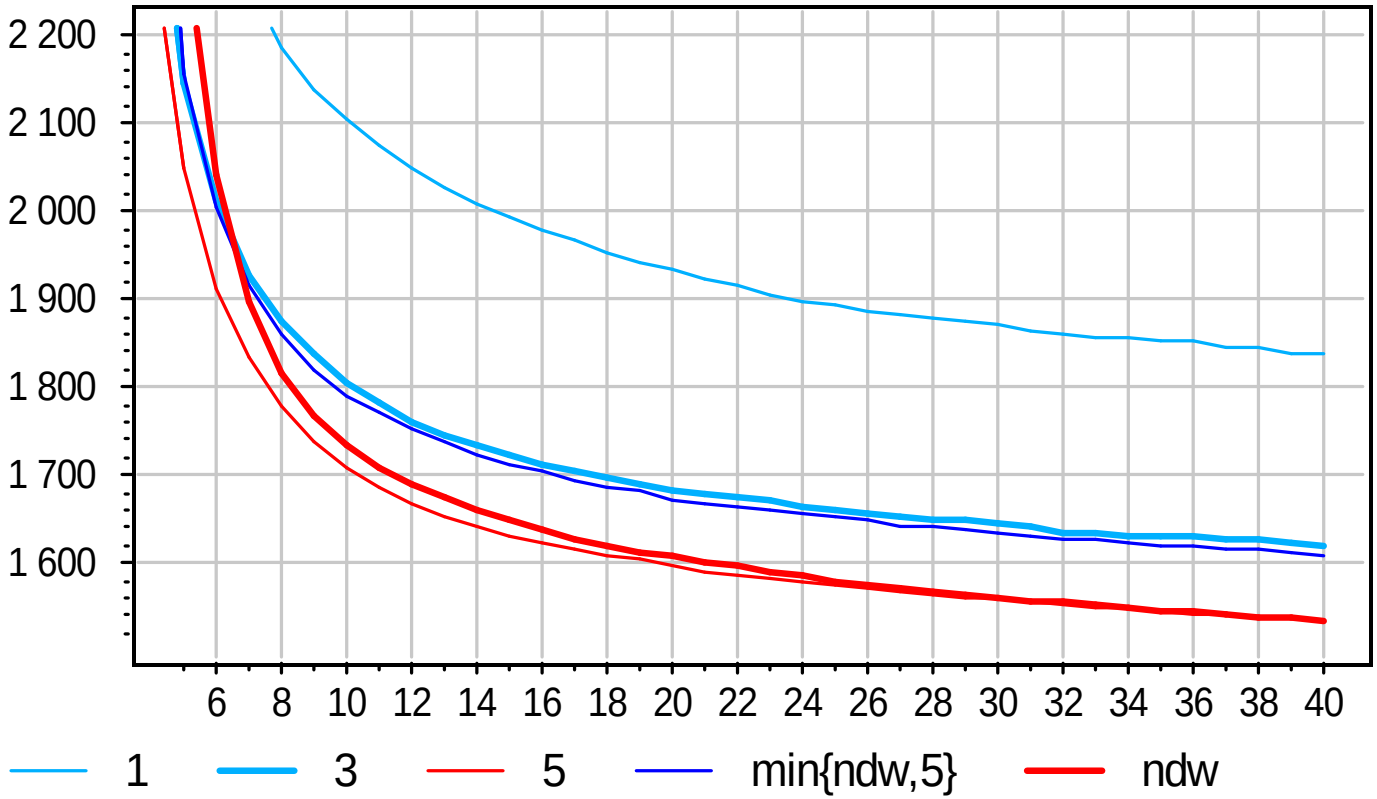
Это противоречит *гипотезе разреженности*:
«употребление слова w в документе d связано с небольшим число тем, чаще всего — с одной».

Идея: для каждой пары (d, w) генерировать s случайных тем t_{dwi}, $i = 1, \ldots, s$, из распределения $p(t|d, w) = H_{dwt}$.

Это эквивалентно замене $p(t|d, w)$ его эмпирической оценкой по сгенерированной случайной выборке длины s:

$$\hat{p}(t|d, w) = \frac{1}{s} \sum_{i=1}^{s} \big[t_{dwi} = t\big].$$

Сколько тем достаточно сэмплировать?



При сэмплировании пяти тем для каждой пары (d, w) перплексия не хуже, чем при сэмплировании n_{dw} тем.
Но одной или трёх тем не достаточно.

Латентное размещение Дирихле
LDA — Latent Dirichlet Allocation [David Blei, 2003]

$p(w|d) = \sum_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$ — модель порождения документа d:

Гипотеза об априорных распределениях Дирихле:

- $\theta_d = (\theta_{td})_{t \in T} \in \mathbb{R}^{|T|}$ — случайные векторы из распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:

$$p(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \quad \sum_t \theta_t = 1;$$

- $\phi_t = (\phi_{wt})_{w \in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:

$$p(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \quad \sum_w \phi_{wt} = 1;$$

Почему именно распределение Дирихле?

Байесовская регуляризация приводит к сглаживанию частотных оценок условных вероятностей:

$$\phi_{wt} := \frac{\hat{n}_{wt} + \beta_w}{\hat{n}_t + \beta_0}; \qquad \theta_{td} := \frac{\hat{n}_{dt} + \alpha_t}{\hat{n}_d + \alpha_0};$$

- При этом устраняются основные недостатки PLSA.
- В экспериментах перплексия LDA меньше, чем у PLSA.
- Распределение Дирихле позволяет описывать кластерную структуру множества мультиномиальных распределений.
- Гиперпараметры α_t, β_w позволяют (в некоторой степени) управлять разреженностью θ_{td}, ϕ_{wt}.

David Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Алгоритмы обучения параметров модели LDA

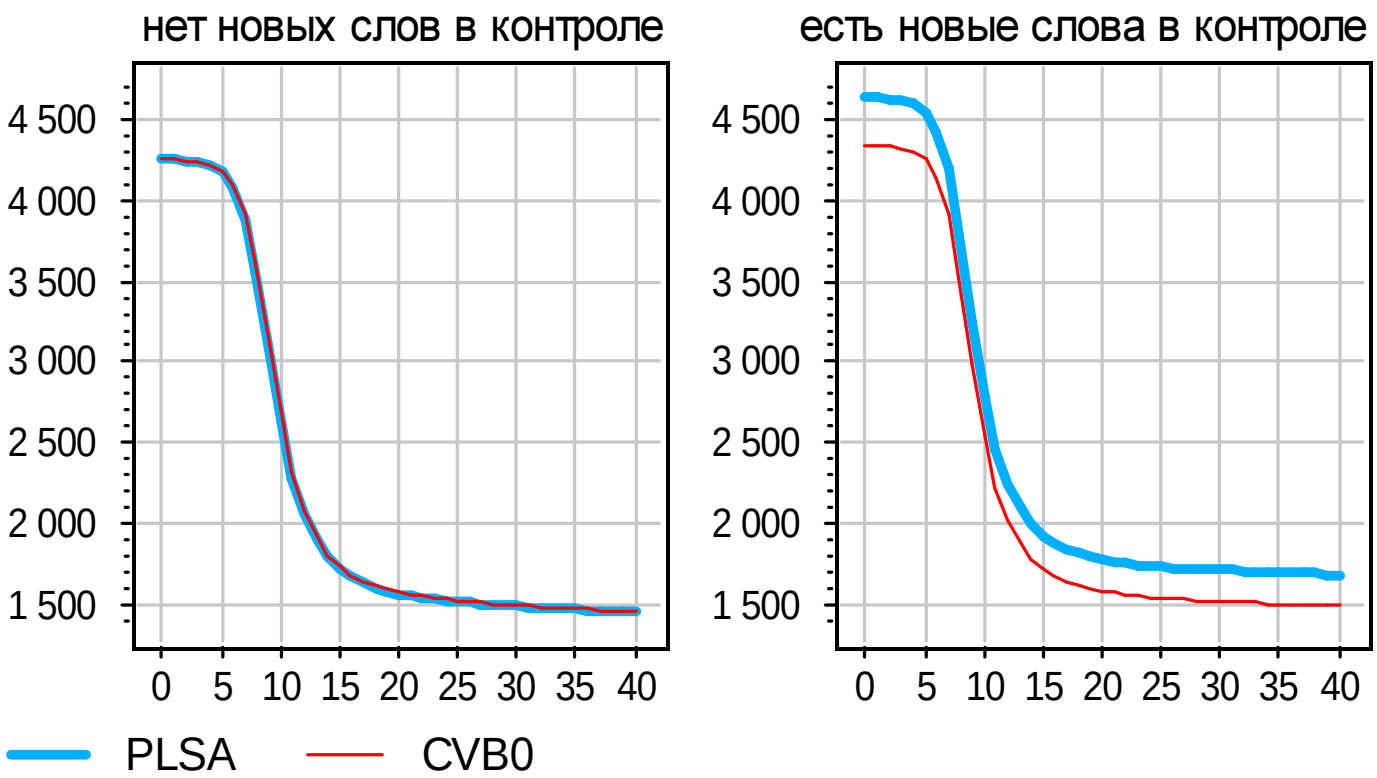
- Сэмплирование Гиббса (GS — Gibbs Sampling)
  можно рассматривать как специальный случай GEM:
  — обновление по каждой словопозиции (n_{dw} раз);
  — сэмплирование 1 темы для каждой словопозиции;
  — регуляризация с гиперпараметрами α, β;

  Griffiths, Steyvers. Finding scientific topics // Proceedings of the National Academy of Sciences. USA, 2004. — Vol. 101. — Pp. 5228–5235.
- CVB — Collapsed Variational Bayesian inference
  можно рассматривать как специальный случай GEM:
  регуляризация без сэмплирования.

  Teh, Newman, Wellingm. A collapsed variational Bayesian inference algorithm for Latent Dirichlet Allocation // Advances in Neural Information Processing Systems (NIPS). Cambridge, MA, MIT Press, 2006.

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Int'l conf. on Uncertainty in Artificial Intelligence, 2009.

Регуляризация решает проблему новых слов, а не переобучения



PLSA без регуляризации, CVB0 с регуляризацией.

Вывод: регуляризация даёт преимущество только когда в контроле есть новые термины.

Недостатки LDA

- Число параметров Φ, Θ в PLSA и LDA одинаково;
  в LDA необходимо оценивать ещё и гиперпараметры α, β.
- Разреженностью можно управлять с помощью α, β, т.к.

$$\mathrm{E}\theta_{td} = \frac{\alpha_t}{\alpha_0}; \qquad \mathrm{E}\phi_{wt} = \frac{\beta_w}{\beta_0};$$

  при этом θ_{td}, ϕ_{wt} могут быть близки, но не равны 0.
- Сглаженные байесовские оценки являются смещёнными.
- Не ясно, как инициализировать β_w для новых терминов.
  Обычно β_w берут равными, но это очень грубая оценка.

Вывод: байесовская регуляризация тематических моделей не решает проблемы переобучения и обеспечения разреженности и плохо решает проблему оценивания новых слов.

Робастная вероятностная тематическая модель SWB — Special Words with Background [Steyvers et al. 2006]

Гипотеза: каждое употребление термина в документе объясняется либо темой, либо специфично для данного документа (шум), либо это общеупотребительный термин (фон).

Модель смеси тематической, шумовой и фоновой компонент:

$$p(w|d) = \frac{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}{1+\gamma+\varepsilon}; \qquad Z_{dw} = \sum_{t\in T}\phi_{wt}\theta_{td},$$

$\pi_{dw} \equiv p_{\text{ш}}(w|d)$ — *шумовая* компонента, γ — параметр;
$\pi_w \equiv p_{\text{ф}}(w)$ — *фоновая* компонента, ε — параметр.

Требуется найти ϕ_{wt}, θ_{td}, π_{dw}, π_w для всех $d\in D$, $w\in W$, $t\in T$.

Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems, MIT Press, 2006. — Vol. 19. — Pp. 241–248.

EM-алгоритм для робастной модели

E-шаг: вероятности тем, фона и шума для каждого (d, w):

$$H_{dwt} = \frac{\phi_{wt}\theta_{td}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}, \quad t \in T;$$

$$H_{dw} = \frac{\gamma\pi_{dw}}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}; \quad H'_{dw} = \frac{\varepsilon\pi_w}{Z_{dw} + \gamma\pi_{dw} + \varepsilon\pi_w}.$$

M-шаг: решение задачи максимизации правдоподобия

ϕ_{wt}, θ_{td} — вычисляются по формулам PLSA;

$$\pi_w = \frac{\nu'_w}{\nu'}; \quad \nu'_w = \sum_{d \in D} n_{dw} H'_{dw}; \quad \nu' = \sum_{w \in W} \nu'_w;$$

$$\pi_{dw} = \frac{n_{dw} H_{dw}}{\nu_d}; \quad \nu_d = \sum_{w \in d} \pi_{dw};$$

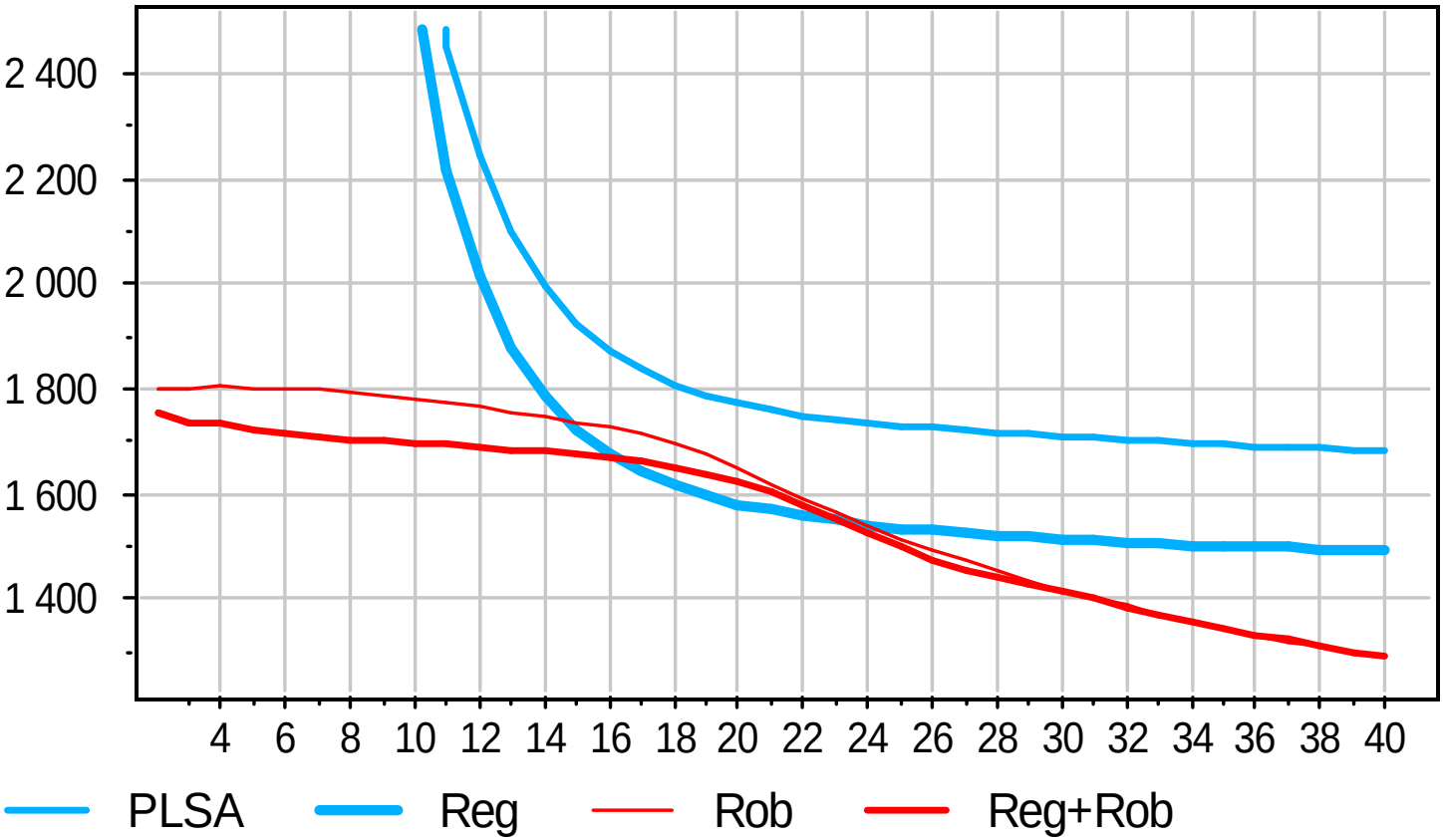
Альтернативный способ оценивания π_{dw} на M-шаге

В робастной модели возможно аналитическое выражение π_{dw} через остальные переменные без вычисления H_{dw},
назовём его *аддитивным M-шагом для шумовой компоненты*:

$$\pi_{dw} = \left(\frac{n_{dw}}{\nu_d} - \frac{Z_{dw} + \varepsilon \pi_w}{\gamma} \right)_+ ,$$

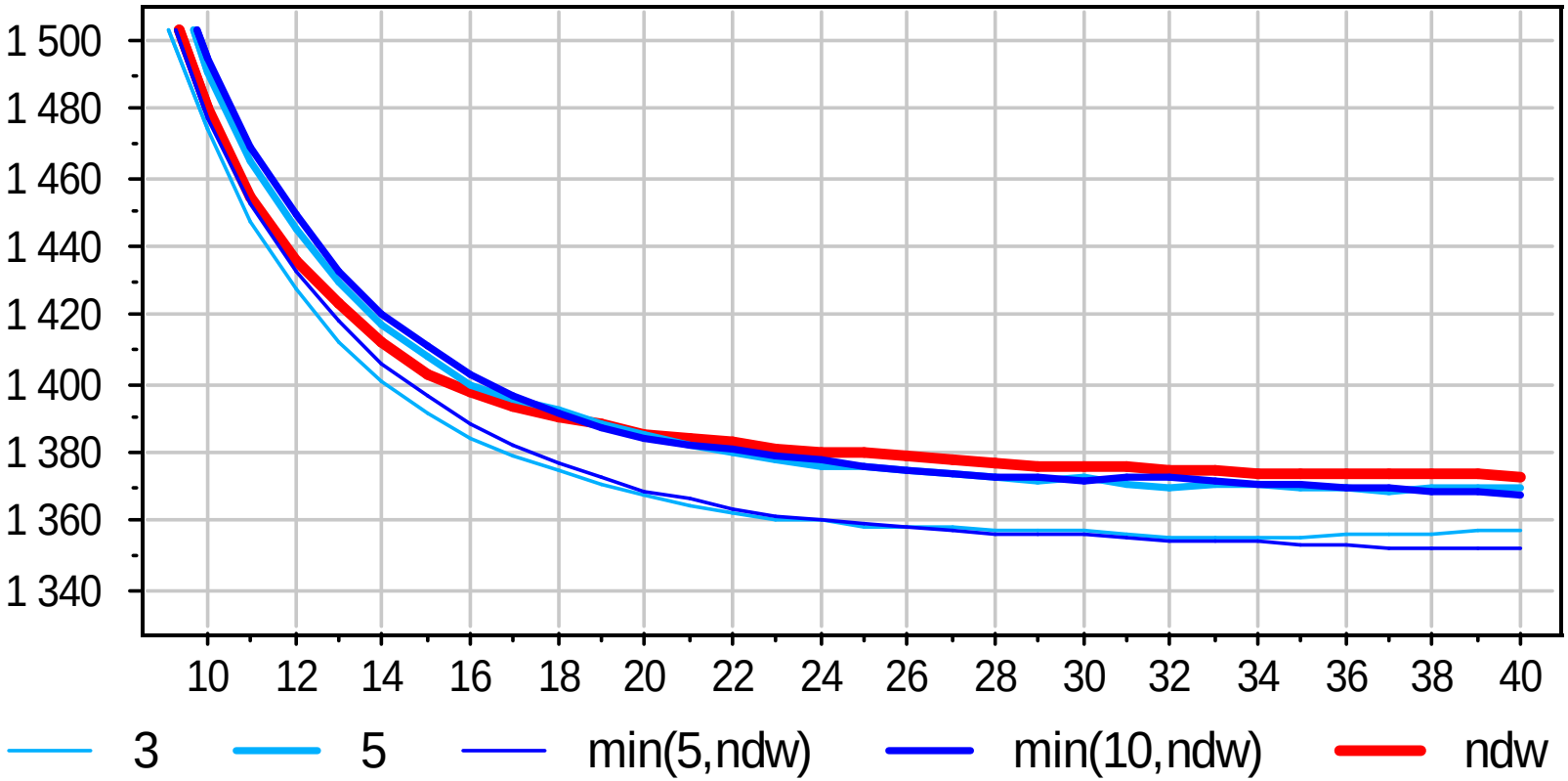
Таким образом, если термин w в документе d встречается существенно чаще, чем предсказывают тематическая и фоновая компоненты модели, то его появление объясняется особенностями данного документа, и тогда $\pi_{dw} > 0$.

Робастная модель не нуждается в регуляризации



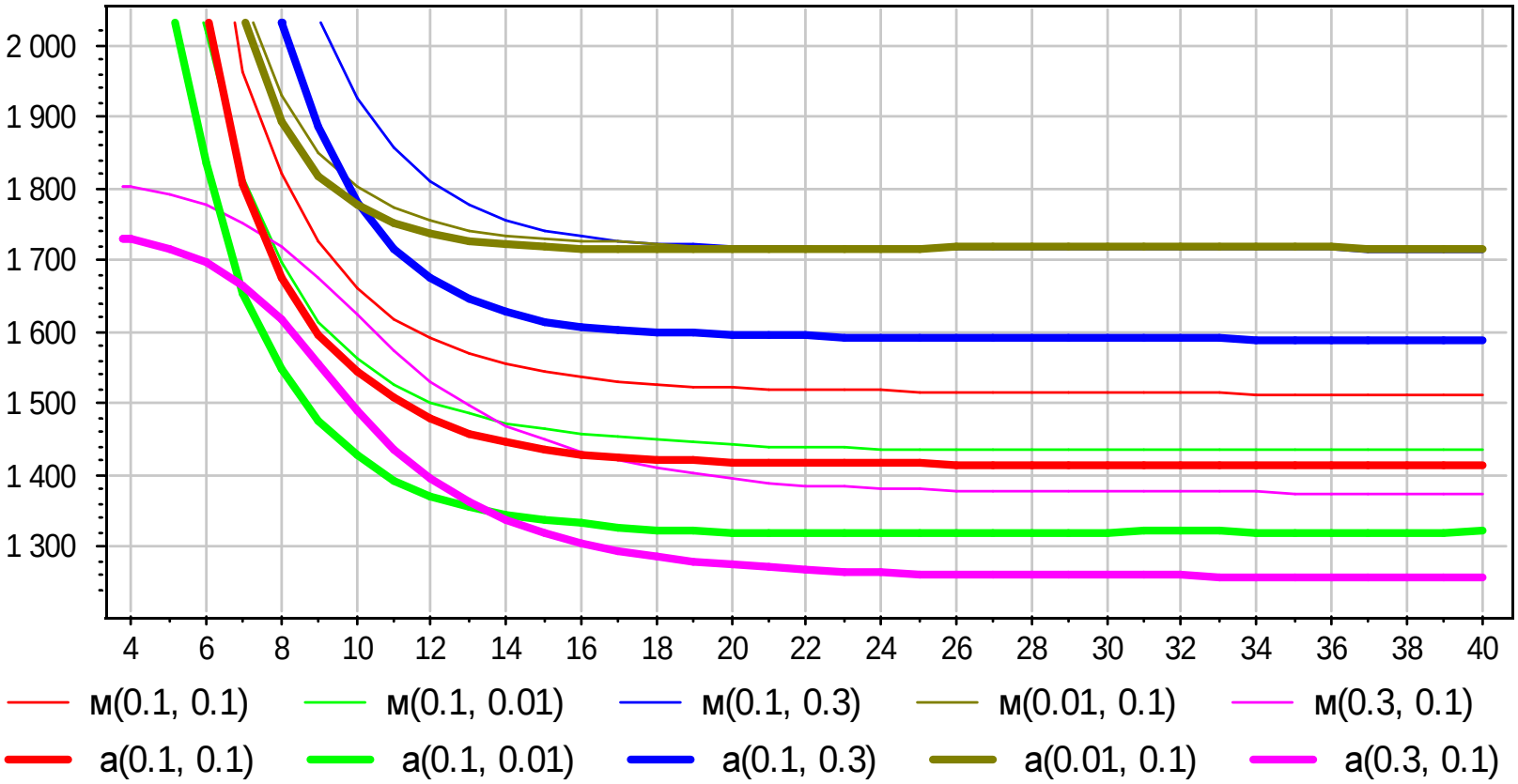
Робастность сильнее уменьшает перплексию PLSA, чем регуляризация. Регуляризация не улучает робастную модель.

Экономное сэмплирование для робастной модели



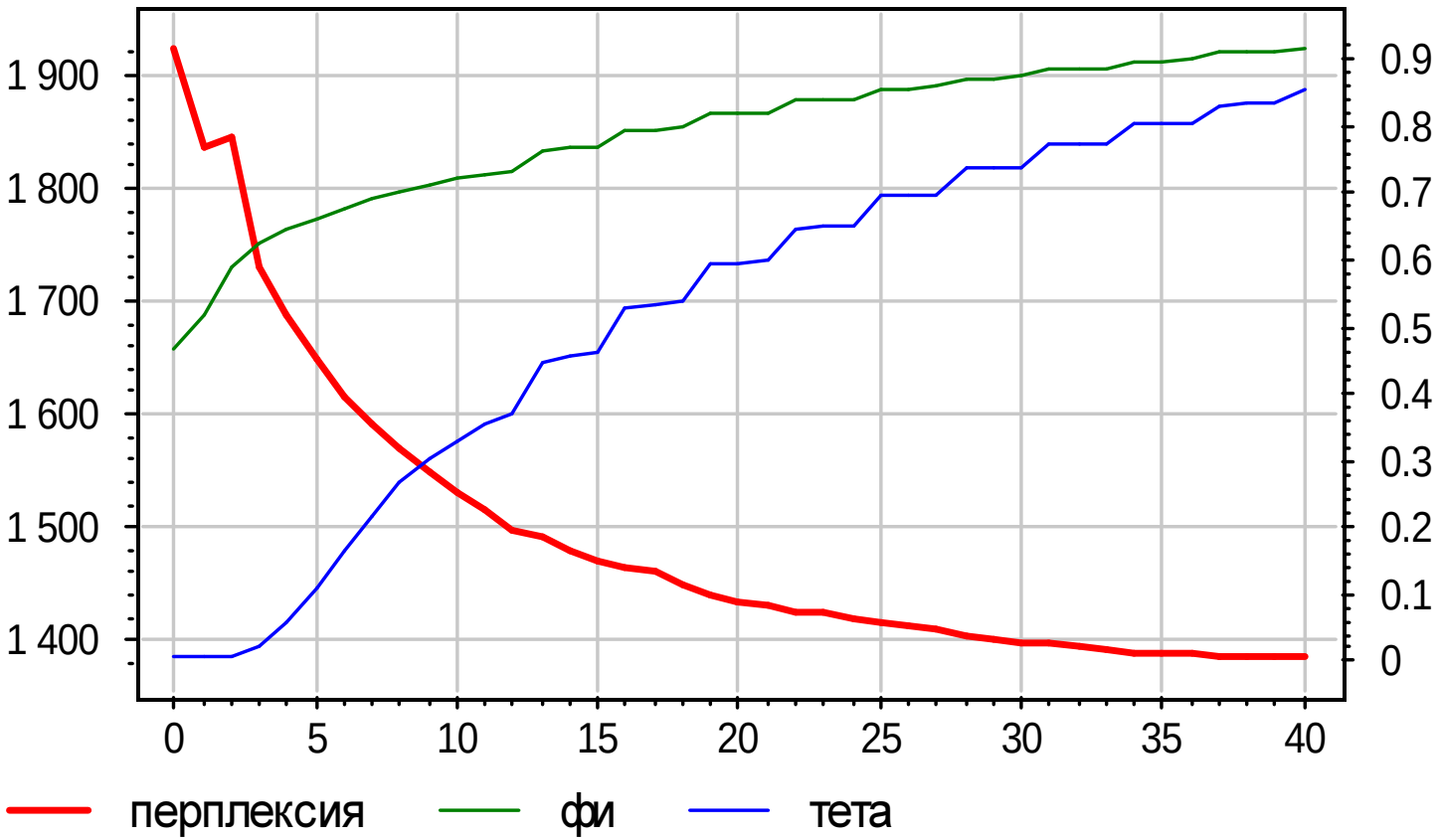
Робастная модель менее чувствительна к выбору числа сэмплируемых тем. Оптимум при $\min\{5, n_{dw}\}$.

Аддитивный и мультипликативный M-шаг



Аддитивный M-шаг a(ε, γ) лучше мультипликативного м(ε, γ).
Перплексия чувствительна к выбору параметров (ε, γ).
Для данной коллекции оптимум при $(\varepsilon, \gamma) = (0.3, 0.1)$.

Робастная модель допускает принудительное разреживание



В процессе разреживания доля нулевых ϕ_{wt} и θ_{td} (отложена по правой оси) увеличивается при монотонном уменьшении перплексии.

Преимущества робастной модели

- Не требуется регуляризация, следовательно
  — используются только несмещённыне оценки,
  — не надо настраивать гиперпараметры.
- Перплексия существенно лучше, чем у LDA.
- Новые слова естественно воспринимаются как шум,
  пока по ним не наберётся выборка,
  достаточная для определения тематики.
- Параметр γ и эвристика принудительного разреживания
  вместе позволяют управлять разреженностью.

Основные результаты

- Предложен GEM-алгоритм обучения тематических моделей, допускающий *в произвольных сочетаниях* эвристики
  — частого обновления параметров;
  — сэмплирования (в том числе экономного);
  — регуляризации;
  — робастности относительно шума;
  — робастности относительно фона.
- PLSA, LDA, SWB реализуются в нём как частные случаи.
- Показано, что робастные модели не нуждаются в регуляризации.
- Показано, что принудительное разреживание не портит робастную модель (в отличие от PLSA и LDA).

Робастность и разреженность необходимы для построения тематических моделей сверхбольших коллекций документов.

Вопросы. . .

Воронцов Константин Вячеславович
vokov@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Тематическое моделирование