Mach Learn (2015) 101:303–323
DOI 10.1007/s10994-014-5476-6

Additive regularization of topic models

Konstantin Vorontsov · Anna Potapenko

Received: 22 January 2014 / Accepted: 18 November 2014 / Published online: 10 December 2014

Abstract Probabilistic topic modeling of text collections has been recently developed mainly within the framework of graphical models and Bayesian inference. In this paper we introduce an alternative semi-probabilistic approach, which we call *additive regularization of topic models* (ARTM). Instead of building a purely probabilistic generative model of text we regularize an ill-posed problem of stochastic matrix factorization by maximizing a weighted sum of the log-likelihood and additional criteria. This approach enables us to combine probabilistic assumptions with linguistic and problem-specific requirements in a single multi-objective topic model. In the theoretical part of the work we derive the regularized EM-algorithm and provide a pool of regularizers, which can be applied together in any combination. We show that many models previously developed within Bayesian framework can be inferred easier within ARTM and in some cases generalized. In the experimental part we show that a combination of sparsing, smoothing, and decorrelation improves several quality measures at once with almost no loss of the likelihood.

Keywords Probabilistic topic modeling · Regularization of ill-posed problems · Probabilistic latent sematic analysis · Latent Dirichlet allocation · EM-algorithm

Editors: Vadim Strijov, Richard Weber, Gerhard-Wilhelm Weber, and Süreyya Ozogur Akyüz.

K. Vorontsov (✉)
Department of Intelligent Systems at Dorodnicyn Computing Centre of RAS,
Institute of Physics and Technology, Moscow, Russia
e-mail: vokov@forecsys.ru

A. Potapenko
Computer Science Department, The Higher School of Economics, Moscow, Russia
e-mail: anya_potapenko@mail.ru

1 Introduction

Topic modeling is a rapidly developing branch of statistical text analysis (Blei 2012). A probabilistic topic model of a text collection defines each topic by a multinomial distribution over words, and then describes each document with a multinomial distribution over topics. Such representation reveals a hidden thematic structure of the collection and promotes the usage of topic models in information retrieval, classification, categorization, summarization and segmentation of texts.

Latent Dirichlet allocation (LDA) (Blei et al. 2003) is the most popular probabilistic topic model. LDA is a two-level Bayesian generative model, in which topic distributions over words and document distributions over topics are generated from prior Dirichlet distributions. This assumption reduces model complexity and facilitates Bayesian inference due to the conjugacy of Dirichlet and multinomial distributions.

Hundreds of LDA extensions have been developed recently to model natural language phenomena and to incorporate additional information about authors, time, labels, categories, citations, links, etc., (Daud et al. 2010).

Nevertheless, building combined and multi-objective topic models remains a difficult problem in Bayesian approach because of a complicated inference in the case of a non-conjugate prior. This open issue is little discussed in the literature. An evolutionary approach has been proposed recently (Khalifa et al. 2013), but it seems to be computationally infeasible for large text collections.

Another difficulty is that Dirichlet prior conflicts with natural assumptions of sparsity. A document usually contains a small number of topics, and a topic usually consists of a small number of domain-specific terms. Therefore, most words and topics must have zero probabilities. Sparsity helps to save memory and time in modeling large text collections. However, Bayesian approaches to sparsing (Shashanka et al. 2008; Wang and Blei 2009; Larsson and Ugander 2011; Eisenstein et al. 2011; Chien and Chang 2013) suffer from an internal contradiction with Dirichlet prior, which can not produce vectors with zero elements.

To address the above problems we introduce a non-Bayesian semi-probabilistic approach—*Additive Regularization of Topic Models* (ARTM). Learning a topic model from a document collection is an ill-posed problem of approximate stochastic matrix factorization, which has an infinite set of solutions. To choose a better solution, we add regularization penalty terms to the log-likelihood. Any problem-oriented regularizers or their linear combination may be used instead of Dirichlet prior or together with it. The idea of ARTM is inspired by Tikhonov's regularization of ill-posed inverse problems (Tikhonov and Arsenin 1977).

Additive regularization differs from Bayesian approach in several aspects.

Firstly, we do not aim to build a fully generative probabilistic model of text. Many requirements for a topic model can be more naturally formalized in terms of optimization criteria rather than prior distributions. Regularizers may have no probabilistic interpretation at all. The structure of regularized models is so straightforward that their representation and explication in terms of graphical models is no longer needed. Thus, ARTM falls into the trend of avoiding excessive probabilistic assumptions in natural language processing.

Secondly, we use the regularized expectation–maximization (EM) algorithm instead of more complicated Bayesian inference. We do not use conjugate priors, integrations, and variational approximations. Despite these fundamental differences both approaches often result in the same or very similar learning algorithms, but in ARTM the inference is much shorter.

Thirdly, ARTM considerably simplifies both design and inference of multi-objective topic models. At the design stage we formalize each requirement for the model in a form of a

regularizer—a criterion to be maximized. At the inference stage we simply differentiate each regularizer with respect to the model parameters.

ARTM also differs from previous regularization techniques each designed for a particular regularizer such as KL-divergence, Dirichlet prior, L_1 or L_2 penalty terms (Si and Jin 2005; Chien and Wu 2008; Wang et al. 2011; Larsson and Ugander 2011). ARTM is not an incremental improvement of a particular topic model, but a new instrument for building and combining topic models much easier than in the state-of-the-art Bayesian approach.

The aim of the paper is to introduce a new regularization framework for topic modeling and to provide an initial pool of useful regularizers.

The rest of the paper is organized as follows.

In Sect. 2 we describe probabilistic latent semantic analysis (PLSA) model, the historical predecessor of LDA. We introduce the EM-algorithm from optimizational point of view. Then we show experimentally on synthetic data that both PLSA and LDA give non-unique and unstable solutions. Further we use PLSA as a more appropriate base for a stronger problem-oriented regularization.

In Sect. 3 we introduce the ARTM approach and prove general equations for regularized EM-algorithm. It is a major theoretical contribution of the paper.

In Sect. 4 we work out a pool of regularizers by revising known topic models. We propose an alternative interpretation of LDA as a regularizer that minimizes Kullback–Leibler divergence with a fixed multinomial distribution. Then we consider regularizers for smoothing, sparsing, semi-supervised learning, topic correlation and decorrelation, topic coherence maximization, documents linking, and document classification. Most of them require tedious calculations within Bayesian approach, whereas ARTM leads to similar results "in one line".

In Sect. 5 we combine three regularizers from our pool to build a highly sparse and well interpretable topic model. We propose to monitor many quality measures during EM-iterations to choose the regularization path empirically for a multi-objective topic model. In our experiment we measure sparsity, kernel size, coherence, purity, and contrast of the topics. We show that ARTM improves all measures at once almost without any loss of the hold-out perplexity.

In Sect. 6 we discuss advantages and limitations of ARTM.

2 Topic models PLSA and LDA

Let D denote a set (collection) of texts and W denote a set (vocabulary) of all terms from these texts. Each term can represent a single word as well as a key phrase. Each document $d \in D$ is a sequence of n_d terms $(w_1, \ldots, w_{n_d})$ from the vocabulary W. Each term might appear multiple times in the same document.

Assume that each term occurrence in each document refers to some latent topic from a finite set of topics T. Text collection is considered to be a sample of triples (w_i, d_i, t_i), $i = 1, \ldots, n$ drawn independently from a discrete distribution $p(w, d, t)$ over a finite space $W \times D \times T$. Term w and document d are observable variables, while topic t is a *latent* (hidden) variable. Following the "bag of words" model, we represent each document by a subset of terms $d \subset W$ and the corresponding integers n_{dw}, which count how many times the term w appears in the document d.

Conditional independence is an assumption that each topic generates terms regardless of the document: $p(w \mid t) = p(w \mid d, t)$. According to the law of total probability and the assumption of conditional independence

$$p(w \mid d) = \sum_{t \in T} p(t \mid d)\, p(w \mid t). \tag{1}$$

The probabilistic model (1) describes how the collection D is generated from the known distributions $p(t \mid d)$ and $p(w \mid t)$. Learning a topic model is an inverse problem: to find distributions $p(t \mid d)$ and $p(w \mid t)$ given a collection D. This problem is equivalent to finding an approximate representation of counter matrix

$$F = \big(\hat{p}_{wd}\big)_{W \times D}, \quad \hat{p}_{wd} = \hat{p}(w \mid d) = \tfrac{n_{dw}}{n_d}, \tag{2}$$

as a product $F \approx \Phi\Theta$ of two unknown matrices—the matrix Φ of *term probabilities for the topics* and the matrix Θ of *topic probabilities for the documents*:

$$\begin{aligned} &\Phi = (\phi_{wt})_{W \times T}, \quad \phi_{wt} = p(w \mid t), \quad \phi_t = (\phi_{wt})_{w \in W}; \\ &\Theta = (\theta_{td})_{T \times D}, \quad \theta_{td} = p(t \mid d), \quad \theta_d = (\theta_{td})_{t \in T}. \end{aligned} \tag{3}$$

Matrices F, Φ and Θ are *stochastic*, that is, they have non-negative and normalized columns representing discrete distributions. Usually the number of topics $|T|$ is much smaller than the collection size $|D|$ and the vocabulary size $|W|$.

In *probabilistic latent semantic analysis* PLSA (Hofmann 1999) the topic model (1) is learned by log-likelihood maximization with linear constrains:

$$L(\Phi, \Theta) = \ln \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Phi, \Theta}; \tag{4}$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \tag{5}$$

Theorem 1 *The stationary point of the optimization problem* (4), (5) *satisfies the system of equations with auxiliary variables* p_{tdw}, n_{wt}, n_{td}, n_t, n_d

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}; \tag{6}$$

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}, \qquad n_t = \sum_{w \in W} n_{wt}; \tag{7}$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}, \qquad n_d = \sum_{t \in T} n_{td}. \tag{8}$$

This statement follows from Karush–Kuhn–Tucker (KKT) conditions. We will prove a more general theorem in the sequel. The system of Eqs. (6)–(8) can be solved by various numerical methods. Particularly, the simple-iteration method is equivalent to the EM algorithm, which is typically used in practice.

EM algorithm repeats two steps in a loop.

The *expectation step* or E-step (6) can be understood as the Bayes' rule for the probability distribution $p(t \mid d, w)$:

$$p_{tdw} = p(t \mid d, w) = \frac{p(w, t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}. \tag{9}$$

The value $n_{tdw} = n_{dw} p_{tdw}$ estimates how many times the term w appears in the document d with relation to the topic t.

The *maximization step* or M-step (7), (8) can therefore be interpreted as frequency estimates for the conditional probabilities ϕ_{wt} and θ_{td}.

Algorithm 2.1: The rational EM-algorithm for PLSA

Input: document collection D, number of topics $|T|$;
Output: Φ, Θ;

initialize vectors ϕ_t, θ_d randomly;
repeat
  zeroize n_{wt}, n_{td}, n_t, n_d for all $d \in D,\ w \in W,\ t \in T$;
  forall $d \in D,\ w \in d$ **do**
    $Z := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    forall $t \in T:\ \phi_{wt}\theta_{td} > 0$ **do**
      increase n_{wt}, n_{td}, n_t, n_d by $\delta = n_{dw}\phi_{wt}\theta_{td}/Z$;
  $\phi_{wt} := n_{wt}/n_t$ for all $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ for all $d \in D,\ t \in T$;
until Φ and Θ converge;

Algorithm 2.1 reorganizes EM iterations by incorporating the E-step inside the M-step. Thus it avoids storage of a three-dimensional array p_{tdw}. Each EM iteration is a run through the entire collection.

Equations (6)–(8) can be rewritten in a shorter notation by omitting normalization and using the proportionality sign: $p_{tdw} \propto \phi_{wt}\theta_{td};\ \phi_{wt} \propto n_{wt};\ \theta_{td} \propto n_{td}$.

In *latent Dirichlet allocation* (LDA) parameters Φ, Θ are constrained by an assumption that vectors ϕ_t and θ_d are drawn from Dirichlet distributions with hyperparameters $\beta = (\beta_w)_{w \in W}$ and $\alpha = (\alpha_t)_{t \in T}$ respectively (Blei et al. 2003). Learning algorithms for LDA generally fall into two categories—sampling-based algorithms (Steyvers and Griffiths 2004) and variational algorithms (Teh et al. 2006). In Gibbs Sampling (LDA-GS) a topic t is sampled from the probability distribution $p(t \mid d, w)$ for each term occurrence $w = w_i$, then counters n_{wt}, n_{td}, n_t, n_t are increased by 1. Learning algorithms for LDA can also be considered as *EM-like algorithms* with modified M-step (Asuncion et al. 2009). The following is the most simple and frequently used modification:

$$\phi_{wt} \propto n_{wt} + \beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_t. \tag{10}$$

It is generally recognized since the work of Blei et al. (2003) that LDA is less subjected to overfitting than PLSA. Nevertheless, recent experiments show that the performance of PLSA and LDA differs insignificantly on large text collections (Masada et al. 2008; Wu et al. 2010; Lu et al. 2011). The reason is that the optimal values of hyperparameters β_w and α_t are usually close to zero (Wallach et al. 2009). Therefore they affect only small values n_{wt} and n_{td} corresponding to the rare terms of topics and rare topics of documents. Robust variants of PLSA and LDA models describe rare terms by a separate model component and have nearly the same performance (Potapenko and Vorontsov 2013). This means that LDA reduces overfitting only for insignificantly rare terms and topics. Thus overfitting does not seems to be such a serious problem for probabilistic topic models.

In contrast, the non-uniqueness, which causes the instability of the solution, is a serious problem. The likelihood (4) depends on the product $\Phi\Theta$, which is defined up to a linear transformation: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, where $\Phi' = \Phi S$ and $\Theta' = S^{-1}\Theta$ are stochastic matrices. The transformation S is not controlled by EM-like algorithms and may depend on random initialization.

We performed the following experiment on the synthetic data in order to assess the ability of PLSA and LDA to restore true matrices Φ, Θ. The collection was generated with the parameters $|W| = 1{,}000$, $|D| = 500$, $|T| = 30$, the lengths of the documents $n_d \in [100, 600]$

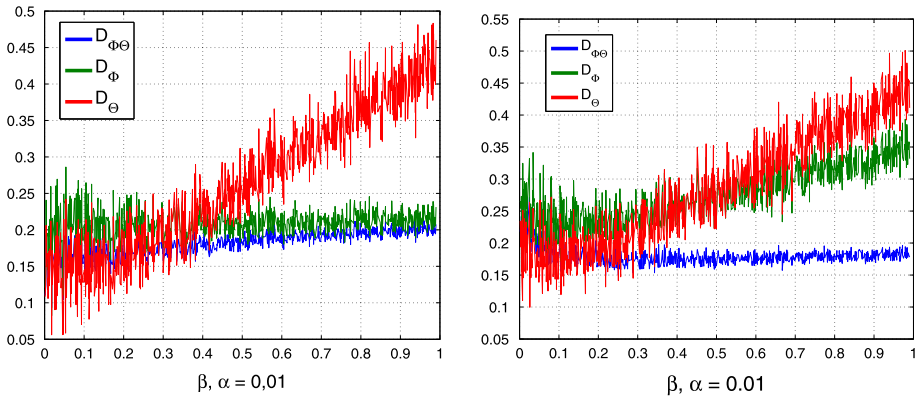

Fig. 1 Errors in restoring the matrices Φ, Θ and $\Phi\Theta$ over hyperparameter β while $\alpha = 0.01$ is fixed for LDA Gibbs sampling (*left chart*) and PLSA-EM (*right chart*)

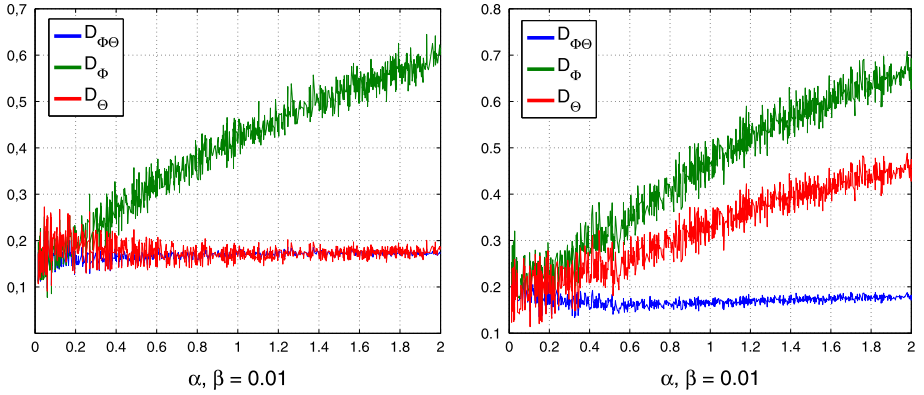

Fig. 2 Errors in restoring the matrices Φ, Θ and $\Phi\Theta$ over hyperparameter α while $\beta = 0.01$ is fixed for LDA Gibbs Sampling (*left chart*) and PLSA-EM (*right chart*)

were chosen randomly. Columns of the matrices Φ, Θ were drawn from the symmetric Dirichlet distributions with parameters β, α respectively. The differences between the restored distributions $\hat{p}(i \mid j)$ and the synthetic ones $p(i \mid j)$, $j = 1, \ldots, m$ were measured by the average Hellinger distance both for the matrices Φ, Θ and for their product:

$$D_\Phi = H(\hat{\Phi}, \Phi); \quad D_\Theta = H(\hat{\Theta}, \Theta); \quad D_{\Phi\Theta} = H(\hat{\Phi}\hat{\Theta}, \Phi\Theta); \tag{11}$$

$$H(\hat{p}, p) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_i \left(\sqrt{\hat{p}(i \mid j)} - \sqrt{p(i \mid j)} \right)^2}. \tag{12}$$

PLSA and LDA turned out to restore the matrices Φ, Θ much worse than their product, Figs. 1, 2. The error depends on the sparsity of the original matrices Φ, Θ. In our experiments LDA did not perform well even when we used the same hyperparameters α, β for synthetic data generation and for LDA-GS algorithm.

These facts show that the Dirichlet distribution is too weak as a regularizer. More problem-oriented regularizers are needed to formalize additional restrictions on the matrices Φ, Θ and to ensure uniqueness and stability of the solution. Therefore our starting point will be the

PLSA model, free of regularizers, but not the LDA model, even though it is more popular in recent research works.

3 EM-algorithm with additive regularization

Consider r additional objectives $R_i(\Phi, \Theta)$, $i = 1, \ldots, r$, called *regularizers*. To maximize these objectives together with the likelihood (4) consider their linear combination with non-negative *regularization coefficients* τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta), \qquad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}. \tag{13}$$

Topic t is called *regular* if $n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} > 0$ for at least one term $w \in W$. If the reverse inequality holds for all $w \in W$ then topic t is called *overregularized*.

Document d is called *regular* if $n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} > 0$ for at least one topic $t \in T$. If the reverse inequality holds for all $t \in T$ then document d is called *overregularized*.

Theorem 2 *If the function $R(\Phi, \Theta)$ is continuously differentiable and (Φ, Θ) is the local maximum of the problem (13), (5), then for any regular topic t and any regular document d the system of equations holds:*

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}; \tag{14}$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+; \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \tag{15}$$

$$\theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+; \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw}; \tag{16}$$

where $(z)_+ = \max\{z, 0\}$.

Note 1 If a topic t is overregularized then (15) gives $\phi_t = 0$. In this case we have to exclude the topic t from the model. Topic overregularization is a mechanism that can eliminate irrelevant topics and optimize the number of topics.

Note 2 If a document d is overregularized then Eq. (16) gives $\theta_d = 0$. In this case we have to exclude the document d from the model. For example, a document may be too short, or have no relation to the thematics of a given collection.

Note 3 Theorem 1 is the particular case of Theorem 2 at $R(\Phi, \Theta) = 0$.

Proof For the local minimum (Φ, Θ) of the problem (13), (5) the KKT conditions can be written as follows:

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w \mid d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \lambda_{wt}; \quad \lambda_{wt} \geqslant 0; \quad \lambda_{wt}\phi_{wt} = 0, \tag{17}$$

where λ_t and λ_{wt} are KKT multipliers for normalization and nonnegativity constrains respectively. Let us multiply both sides of the first equation by ϕ_{wt}, identify the right-hand side of

(14) and replace it by the left-hand side variable p_{tdw}. Then we apply the definition of n_{wt} from (15):

$$\phi_{wt}\lambda_t = \sum_d n_{dw}\frac{\phi_{wt}\theta_{td}}{p(w \mid d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}. \tag{18}$$

An assumption that $\lambda_t \leqslant 0$ contradicts the regularity condition for the topic t. Then $\lambda_t > 0$, $\phi_{wt} \geqslant 0$. The left-hand side is nonnegative, thus the right-hand side is nonnegative too, consequently,

$$\phi_{wt}\lambda_t = \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+. \tag{19}$$

Let us sum both sides of this equation over all $w \in W$:

$$\lambda_t = \sum_{w\in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+. \tag{20}$$

Finally, we obtain (15) by expressing ϕ_{wt} from (19) and (20).
Equations for θ_{td} are derived analogously thus finalizing the proof. □

The system of Eqs. (14)–(16) defines a regularized EM-algorithm. It keeps E-step (6) and redefines M-step by regularized Eqs. (15), (16). Thus, the EM-algorithm for learning regularized topic models can be implemented by easy modification of any EM-like algorithm at hand. Particularly, in Algorithm 2.1 we are to modify only steps 8 and 9 according to Eqs. (15), (16).

4 Regularization criteria for topic models

In this section we collect a pool of regularizers that can be used in any combination or separately. We revise some of well-known topic models that were originally developed within Bayesian approach. We show that ARTM gives similar or more general results through a much simpler inference based on Theorem 2.

We will intensively use the Kullback–Leibler divergence (relative entropy) to measure the difference between multinomial distributions $(p_i)_{i=1}^n$ and $(q_i)_{i=1}^n$:

$$\mathrm{KL}(p\|q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln\frac{p_i}{q_i}. \tag{21}$$

Recall that the minimization of the KL-divergence is equivalent to maximizing the likelihood of the model distribution q for the empirical distribution p.

Smoothing regularization and LDA Let us minimize the KL-divergence between the distributions ϕ_t and a fixed distribution $\beta = (\beta_w)_{w\in W}$, and the KL-divergence between θ_d and a fixed distribution $\alpha = (\alpha_t)_{t\in T}$:

$$\sum_{t\in T}\mathrm{KL}_w(\beta_w\|\phi_{wt}) \to \min_{\Phi}, \qquad \sum_{d\in D}\mathrm{KL}_t(\alpha_t\|\theta_{td}) \to \min_{\Theta}. \tag{22}$$

After summing these criteria with coefficients β_0, α_0 and removing constants we get the regularizer

$$R(\Phi,\Theta) = \beta_0\sum_{t\in T}\sum_{w\in W}\beta_w \ln\phi_{wt} + \alpha_0\sum_{d\in D}\sum_{t\in T}\alpha_t \ln\theta_{td} \to \max. \tag{23}$$

The regularized M-step (15) and (16) gives equations

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_0\alpha_t, \tag{24}$$

which are exactly the same as the M-step (10) in LDA model with hyperparameter vectors $\beta = \beta_0(\beta_w)_{w\in W}$ and $\alpha = \alpha_0(\alpha_t)_{t\in T}$ of the Dirichlet distributions.

The non-Bayesian interpretation of the smoothing regularization in terms of KL-divergence is simple, natural, and avoids complicated inference.

Sparsing regularization The opposite regularization strategy is to maximize KL-divergence between ϕ_t, θ_d and fixed distributions β, α:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t\in T}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} \to \max. \tag{25}$$

For example, to find a sparse distributions ϕ_{wt} with lower entropy we may choose the uniform distribution $\beta_w = \frac{1}{|W|}$, which is known to have the largest entropy.

The regularized M-step (15) and (16) gives equations that differ from the smoothing equations in the sign of the parameters β, α:

$$\phi_{wt} \propto \big(n_{wt} - \beta_0\beta_w\big)_+, \qquad \theta_{td} \propto \big(n_{td} - \alpha_0\alpha_t\big)_+. \tag{26}$$

The idea of entropy-based sparsing was originally proposed in the dynamic PLSA for video processing to produce sparse distributions of topics over time (Varadarajan et al. 2010). The conflict between Dirichlet prior and sparsing assumption leads to sophisticated sparse LDA models (Shashanka et al. 2008; Wang and Blei 2009; Eisenstein et al. 2011; Larsson and Ugander 2011; Chien and Chang 2013). A simple and natural sparsing becomes possible due to abandoning the Dirichlet prior within ARTM semi-probabilistic regularization framework.

Smoothing regularization for semi-supervised learning Consider a collection, which is partially labeled by experts: each document d from a subset $D_0 \subseteq D$ is associated with a subset of topics $T_d \subset T$, and each topic t from a subset $T_0 \subset T$ is associated with a subset of terms $W_t \subset W$. It is usually expected that labeling information helps to improve the interpretability of topics.

Consider the regularizer that minimizes KL-divergence between ϕ_t, θ_d and uniform distributions $\beta_{wt} = \frac{1}{|W_t|}[w \in W_t]$, $\alpha_{td} = \frac{1}{|T_d|}[t \in T_d]$ respectively:

$$R(\Phi, \Theta) = \beta_0 \sum_{t\in T_0}\sum_{w\in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d\in D_0}\sum_{t\in T} \alpha_{td} \ln \theta_{td} \to \max. \tag{27}$$

The regularized M-step (15) and (16) gives another kind of smoothing:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_{wt}\,[t \in T_0]; \tag{28}$$

$$\theta_{td} \propto n_{td} + \alpha_0\alpha_{td}\,[d \in D_0]. \tag{29}$$

This can be considered as yet another generalization of LDA, in which vectors β, α are different for the respective distributions ϕ_t, θ_d depending on labeled data.

Decorrelation of topics Reducing the overlapping between the topic-word distributions is known to make the learned topics more interpretable (Tan and Ou 2010). A regularizer that minimizes covariance between vectors ϕ_t,

$$R(\Phi) = -\gamma \sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \phi_{wt}\phi_{ws} \to \max, \tag{30}$$

leads to the following equation of the M-step:

$$\phi_{wt} \propto \Big(n_{wt} - \gamma\phi_{wt} \sum_{s\in T\setminus t} \phi_{ws}\Big)_+. \tag{31}$$

From this formula we conclude that for each term w the highest probabilities ϕ_{wt} will increase even further, while small probabilities will decrease from iteration to iteration, and may eventually turn into zeros. Therefore, this regularizer also stimulates sparsity. Besides, it has another useful property, which is to group stop-words into a separate topic (Tan and Ou 2010).

Covariance regularization for documents Sometimes we possess information that some documents are likely to share similar topics. For example, they may fall into the same category or one document may have a reference or a link to the other. Making use of this information in terms of the regularizer, we get:

$$R(\Theta) = \tau \sum_{d,c} n_{dc} \sum_{t\in T} \theta_{td}\theta_{tc} \to \max, \tag{32}$$

where n_{dc} is the weight of the link between documents d and c. A similar model LDA-JS by Dietz et al. (2007) is based on the minimization of Jensen–Shannon divergence between θ_d and θ_c, rather than on the covariance maximization.

According to (16), the equation for θ_{td} in the M-step turns into

$$\theta_{td} \propto n_{td} + \tau\theta_{td} \sum_{c\in D} n_{dc}\theta_{tc}. \tag{33}$$

This is a kind of smoothing regularizer, which adjusts probabilities θ_{td} so that they become closer to θ_{tc} for all documents c, connected with d.

Correlated topic model (CTM) was first introduced by Blei and Lafferty (2007) to find strong correlations between topics. For example, a document about geology is more likely to also be about archeology than genetics.

In CTM the correlation between topics is modeled by an assumption that document vectors θ_d are generated by *logistic normal* prior distribution:

$$\theta_{td} = \frac{\exp(\eta_{td})}{\sum_{s\in T}\exp(\eta_{sd})}; \qquad p(\eta_d \mid \mu, \Sigma) = \frac{\exp\big(-\frac{1}{2}(\eta_d-\mu)^{\mathrm{T}}\Sigma^{-1}(\eta_d-\mu)\big)}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}, \tag{34}$$

where $|T|$-vector μ and $|T|\times|T|$ covariance matrix Σ are parameters of Gaussian distribution. Document vectors $\eta_d \in \mathbb{R}^{|T|}$ are determined by the corresponding vectors θ_d up to an arbitrary document-dependent constant C_d:

$$\eta_{td} = \ln\theta_{td} + C_d. \tag{35}$$

Initially CTM was developed within Bayesian approach, although Bayesian inference is complicated by the fact that the logistic normal distribution is not conjugate to the multinomial. We argue that the very idea of CTM can be alternatively implemented and easier understood within ARTM approach.

In terms of ARTM we define a regularizer as the log-likelihood of the logistic normal model for a sample of the document vectors η_d:

$$R(\Theta) = \tau \sum_{d\in D} \ln p(\eta_d \mid \mu, \Sigma) = -\frac{\tau}{2}\sum_{d\in D}(\eta_d-\mu)^{\mathrm{T}}\Sigma^{-1}(\eta_d-\mu) + \text{const} \to \max. \tag{36}$$

According to (16) the equation for θ_{td} in the M-step turns into

$$\theta_{td} \propto \left(n_{td} - \tau \sum_{s \in T} \tilde{\Sigma}_{ts} \left(\ln \theta_{sd} - \mu_s \right) \right)_+ , \tag{37}$$

where $\Sigma^{-1} = (\tilde{\Sigma}_{ts})_{T \times T}$ is the inverse covariance matrix.

The parameters Σ, μ of Gaussian distribution are assumed to be constant during the iteration. Following the idea of block-coordinate optimization we estimate them after each run through the collection (in Algorithm 2.1 after step 9):

$$\mu = \frac{1}{|D|} \sum_{d \in D} \ln \theta_d ; \tag{38}$$

$$\Sigma = \frac{1}{|D|} \sum_{d \in D} \big(\ln \theta_d - \mu \big) \big(\ln \theta_d - \mu \big)^{\mathrm{T}} . \tag{39}$$

Then we invert the covariance matrix and turn insignificant values $\tilde{\Sigma}_{ts}$ into zeros to get sparse solution and reduce computations in (37). Blei and Lafferty (2007) propose to use lasso regression for this purpose.

Coherence regularization A topic is called *coherent* if its most frequent words typically appear nearby in the documents—either in the training collection, or in some external corpus like Wikipedia. An average topic coherence is considered to be a good interpretability measure of a topic model (Newman et al. 2010b).

Let $C_{wv} = \hat{p}(w \mid v)$ denote an estimate of the co-occurrence of word pairs $(w, v) \in W^2$. Usually, C_{wv} is defined as a portion of the documents that contain both words v and w in a sliding window of ten words.

Let us estimate the conditional probability $p(w \mid t)$ from $\phi_{vt} = p(v \mid t)$ over all coherent words v using the law of total probability:

$$\hat{p}(w \mid t) = \sum_{v \in W \setminus w} C_{wv} \phi_{vt} = \sum_{v \in W \setminus w} \frac{C_{wv} n_{vt}}{n_t} . \tag{40}$$

Consider a regularizer which minimizes the weighted sum of KL-divergences between the empirical distribution $\hat{p}(w \mid t)$ and the model distribution ϕ_{wt}:

$$R(\Phi) = \tau \sum_{t \in T} n_t \sum_{w \in W} \hat{p}(w \mid t) \ln \phi_{wt} \to \max . \tag{41}$$

According to (15) the equation of the M-step turns into

$$\phi_{wt} \propto n_{wt} + \tau \sum_{v \in W \setminus w} C_{wv} n_{vt} . \tag{42}$$

The same formula was derived by Mimno et al. (2011) for LDA model and Gibbs Sampling algorithm, from more complicated reasoning through a generalized Polya urn model and a more complex heuristic estimate for C_{wv}.

Newman et al. (2011) propose yet another regularizer:

$$R(\Phi) = \tau \sum_{t \in T} \ln \sum_{u,v \in W} C_{uv} \phi_{ut} \phi_{vt} \to \max , \tag{43}$$

where $C_{uv} = N_{uv}$ if $\mathrm{PMI}(u, v) > 0$ and $C_{uv} = 0$ otherwise, pointwise mutual information $\mathrm{PMI}(u, v) = \ln \frac{|D| N_{uv}}{N_u N_v}$ depends on document frequencies: N_{uv} is the number of documents

that contain both words u, v in a sliding window of ten words, N_u is the number of documents that contain at least one occurrence of the word u.

Thus we conclude that there is no commonly accepted approach to the coherence optimization in the literature. All approaches that we have found so far can be easily expressed in terms of ARTM without Dirichlet priors.

Document classification Let C be a finite set of classes. Suppose each document d is labeled by a subset of classes $C_d \subset C$. The task is to infer a relationship between classes and topics, to improve a topic model by using labeling information, and to learn a decision rule, which is able to classify new documents. Common discriminative approaches such as SVM or Logistic Regression usually give unsatisfactory results on large text collections with a big number of unbalanced and interdependent classes. Probabilistic topic model can benefit in this situation because it processes all classes simultaneously (Rubin et al. 2012).

There are many examples of document labeling in the literature. Classes may refer to text categories (Rubin et al. 2012; Zhou et al. 2009), authors (Rosen-Zvi et al. 2004), time periods (Cui et al. 2011; Varadarajan et al. 2010), cited documents (Dietz et al. 2007), cited authors (Kataria et al. 2011), users of documents (Wang and Blei 2011). More information about special models can be found in the survey (Daud et al. 2010). All these models fall into several groups and all of them can be easily expressed in terms of ARTM. Below we consider a close analogue of Dependency LDA (Rubin et al. 2012), one of the most general topic models for document classification.

We expand the probability space to the set $D \times W \times T \times C$ and assume that each term w in document d is related to both topic $t \in T$ and class $c \in C$. To classify documents we model distribution $p(c \mid d)$ over classes for each document d. We assume that classes of a document are determined by its topics, then conditional independence assumption $p(c \mid t) = p(c \mid d, t)$ is satisfied. This allows us to express $p(c \mid d)$ in terms of *class probabilities for the topics* $p(c \mid t) = \psi_{ct}$ and topic probabilities for the documents $p(t \mid d) = \theta_{td}$ in the way that is similar to the basic topic model (1):

$$p(c \mid d) = \sum_{t \in T} \psi_{ct} \theta_{td}. \tag{44}$$

Thus we introduce a third stochastic matrix of model parameters $\Psi = (\psi_{ct})_{C \times T}$.

Another conditional independence $p(w, c \mid d) = p(w \mid d)\, p(c \mid d)$ allows to split the log-likelihood into PLSA term $L(\Phi, \Theta)$ as in (4) and a regularization term $Q(\Psi, \Theta)$:

$$\ln \prod_{d \in D} \prod_{w \in d} p(w, c \mid d)^{n_{dw}} = L(\Phi, \Theta) + \tau Q(\Psi, \Theta) \;\to\; \max_{\Phi, \Theta, \Psi}; \tag{45}$$

$$Q(\Psi, \Theta) = \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct} \theta_{td}, \tag{46}$$

where m_{dc} is the empirical frequency of classes in document d. It can be estimated via uniform distribution over classes: $m_{dc} = n_d \frac{[c \in C_d]}{|C_d|}$. The regularization coefficient τ may be set to 1 or it may be used to trade-off the document language model $p(w \mid d)$ and the document classification model $p(c \mid d)$. The regularizer Q can be considered as a minimization of KL-divergence between the probability model of classification $p(c \mid d)$ and the empirical class frequency m_{dc}. The problem (45), (46) can still be solved via the regularized EM-like algorithm due to the following generalization of Theorem 2.

Theorem 3 *If the function $R(\Phi, \Psi, \Theta)$ of stochastic matrices Φ, Ψ, Θ is continuously differentiable and (Φ, Ψ, Θ) is the local maximum of $L(\Phi, \Theta) + \tau Q(\Psi, \Theta) + R(\Phi, \Psi, \Theta)$*

then for any regular topic t and any regular document d the system of equations holds:

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}; \qquad p_{tdc} = \frac{\psi_{ct}\theta_{td}}{\sum_{s\in T}\psi_{cs}\theta_{sd}}; \tag{47}$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+; \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{48}$$

$$\psi_{ct} \propto \left(m_{ct} + \psi_{ct}\frac{\partial R}{\partial \psi_{ct}}\right)_+; \qquad m_{ct} = \sum_{d\in D} m_{dc}p_{tdc}; \tag{49}$$

$$\theta_{td} \propto \left(n_{td} + \tau m_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+; \quad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}; \quad m_{td} = \sum_{c\in C_d} m_{dc}p_{tdc}. \tag{50}$$

We omit the proof, which is analogous to the proof of Theorem 2.

Regularization term $R(\Phi, \Psi, \Theta)$ can include Dirichlet prior for Ψ, as in Dependency LDA, but sparsing seems to be a more natural choice.

Another useful example of R is label regularization.

Label regularization is known to improve multi-label classification for unbalanced classes (Mann and McCallum 2007; Rubin et al. 2012). We encourage the similarity between the model distribution $p(c)$ and the empirical class frequency $\hat{p}_c$ in the training data:

$$R(\Psi) = \xi \sum_{c\in C} \hat{p}_c \ln p(c) \to \max, \quad p(c) = \sum_{t\in T} \psi_{ct}p(t), \quad p(t) = \frac{n_t}{n}, \tag{51}$$

where ξ is the regularization coefficient. The formula for the M-step (49)

$$\psi_{ct} \propto m_{ct} + \xi\hat{p}_c\frac{\psi_{ct}n_t}{\sum_{s\in T}\psi_{cs}n_s} \tag{52}$$

results in smoothing of distributions ψ_{ct} proportionally to the frequencies $\hat{p}_c$.

5 Combining regularizers for sparsing and improving interpretability

Interpretability of a topic is a poorly formalized requirement. Essentially what it means is that, provided with the list of the most frequent terms and the most representative documents of a topic, a human can understand its meaning and give it an appropriate name. The interpretability is an important property for information retrieval, systematization and visualization of text collections.

Most of the existing approaches involve human assessment. Newman et al. (2009) ask experts to assess the usefulness of topics by a 3-point scale. Chang et al. (2009) prepare lists of 10 most frequent words for each topic, intruding one random word into each list. A topic is considered to be interpretable if experts can correctly identify the intrusion word. Human-based approach is important at research stage, but it prohibits a fully automatic construction of the topic model.

Coherence is the most popular automatic measure, which is known to correlate well with human estimates of the interpretability (Newman et al. 2010a, b; Mimno et al. 2011). Coherence measures how often the most probable words of the topic occur nearby in the documents from the underlying collection or from external polythematic collection such as Wikipedia.

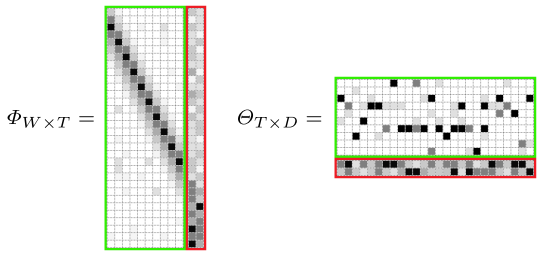

Fig. 3 The example of sparse matrices Φ and Θ with specific and background topics. Background topics are shown as two rightmost columns in Φ and two lowest rows in Θ

In this paper we propose another formalization of interpretability, which also does not require human assessment. We assume that each interpretable topic contains its own *lexical kernel*—a set of specific terms for a particular domain area, which have high probability in this topic, and lower probabilities in other topics. Lexical kernel of the topic should be free of common lexis words, which frequently occur in many documents. Thus, we want to find matrices Φ and Θ with a sparsity structure similar to the one displayed in Fig. 3. To do this we split the set of topics T into two subsets: domain-specific topics S and background topics B.

Domain-specific topic $t \in S$ contains terms of a particular domain area. Domain-specific distributions $p(w \mid t)$ are sparse and weakly correlated. Their corresponding distributions $p(d \mid t)$ are also sparse, because each domain-specific topic occurs in a relatively small number of documents.

Background topic $t \in B$ contains common lexis words. Background distributions $p(w \mid t)$ and $p(d \mid t)$ are smooth, because background words occur in many documents. A topic model with background can be considered as a generalization of robust models, which use only one background distribution (Chemudugunta et al. 2007; Potapenko and Vorontsov 2013).

Combining sparsing, smoothing, and decorrelation To obtain the sparsity sructure of Φ and Θ matrices as shown in Fig. 3, we propose a combination of five regilarizers: smoothing of background topics in matrices Φ and Θ, sparsing of domain-specific topics in matrices Φ and Θ, and decorrelation of domain-specific topics in matrix Φ:

$$\begin{aligned} R(\Phi,\Theta) = &- \beta_0 \sum_{t\in S}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in S} \alpha_t \ln \theta_{td} \\ &+ \beta_1 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_1 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \\ &- \gamma \sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \phi_{wt}\phi_{ws} \to \max . \end{aligned} \tag{53}$$

We use uniform distribution α_t and two types of background distribution β_w: either a uniform distribution, or the term frequency estimates $\beta_w = n_w/n$.

Then we obtain M-step formulas for a combined model from (15) and (16):

$$\phi_{wt} \propto \Bigg(n_{wt} - \beta_0 \underbrace{\beta_w[t\in S]}_{\substack{\text{sparsing}\\ \text{specific}\\ \text{topic}}} + \beta_1 \underbrace{\beta_w[t\in B]}_{\substack{\text{smoothing}\\ \text{background}\\ \text{topic}}} - \underbrace{\gamma\,[t\in S]\,\phi_{wt} \sum_{s\in S\setminus t} \phi_{ws}}_{\text{decorrelation}} \Bigg)_+ ; \tag{54}$$

$$\theta_{td} \propto \Bigg(n_{td} - \alpha_0 \underbrace{\alpha_t[t \in S]}_{\substack{\text{sparsing}\\ \text{specific}\\ \text{topic}}} + \alpha_1 \underbrace{\alpha_t[t \in B]}_{\substack{\text{smoothing}\\ \text{background}\\ \text{topic}}} \Bigg)_+ . \tag{55}$$

Regularization trajectory A linear combination of multiple regularizers R_i depends on a vector of regularization coefficients $\boldsymbol{\tau} = (\tau_i)_{i=1}^r$, which is hard to optimize. A similar problem has been efficiently solved in ElasticNet with a regularization path technique specially developed for a combination of L_1 and L_2 regularization (Friedman et al. 2010). In topic modeling a much larger variety of regularizers is used. Extremely large coefficient may lead to a conflict with other regularizers, to a slower convergence, or to a degeneration of the model. Conversely, extremely small coefficient actually disables the regularization. According to the theory of regularization of ill-posed inverse problems (Tikhonov and Arsenin 1977) we must reduce the regularization coefficient down to zero during the iterations, in order to achieve a correct regularized solution. Optimizing the convergence rate is usually task-dependent and should be controlled manually in the experiment.

Then we define the *regularization trajectory* as a multidimensional vector $\boldsymbol{\tau}$, which is a function of the number of iteration and, possibly, of the model quality measures. In our experiments we choose the regularization trajectory by analyzing experimentally how the change of regularization coefficients affects quality measures of the model during iterations.

Quality measures Learning a topic model from a text collection can be considered as a constrained multi-criteria optimization problem. Therefore, the quality of a topic model should also be measured by a set of criteria. Below we describe a set of quality measures that we use in our experiments.

The accuracy of a topic model $p(w \mid d)$ on the collection D is commonly evaluated in terms of *perplexity*, which is closely related to the likelihood (the lower perplexity is, the better):

$$\mathscr{P}(D, p) = \exp\Big(-\frac{1}{n} L(\Phi, \Theta)\Big) = \exp\Big(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w \mid d)\Big). \tag{56}$$

The *hold-out perplexity* $\mathscr{P}(D', p_D)$ of the model p_D trained on the collection D is evaluated on the test set of documents D' not intersecting D. In our experiments we split the collection in proportion $|D| : |D'| = 9 : 1$. Each document d from the test set is further randomly split into two halves: the first one is used to estimate parameters θ_d, and the second one is used in the perplexity evaluation. The terms in the second halves that did not appear in D are ignored. Parameters ϕ_t are estimated from the training set D.

The *sparsity* of a model is measured by the ratio of zero elements in matrices Φ and Θ over domain-specific topics S.

The *background ratio* is a ratio of background terms over the collection:

$$\mathscr{B} = \frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} n_{dw} p(t \mid d, w). \tag{57}$$

It takes value from 0 to 1. If $\mathscr{B}$ is close to 0 then the model does not eliminate common lexis from domain-specific topics. If $\mathscr{B}$ is close to 1 then the model is degenerated, possibly due to excessive sparsing.

We define the *lexical kernel* W_t of a topic t as a set of terms that distinguish the topic t from the other topics: $W_t = \{w\colon p(t \mid w) > \delta\}$. In our experiments we set $\delta = 0.25$. Then we define a set of measures, which characterize the conformity of the matrix Φ with the sparse structure shown in Fig. 3:

kernel size $\ker_t = |W_t|$, the reasonable values for it are about $\frac{|W|}{|T|}$;
purity $\mathrm{pur}_t = \sum_{w \in W_t} p(w \mid t)$, the higher the better;
contrast $\mathrm{con}_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t \mid w)$, the higher the better.

The *coherence of a topic* t is defined as the pointwise mutual information averaged over all word pairs from the top-k most probable words of the topic t:

$$\mathscr{C}_t^k = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i}^{k} \mathrm{PMI}(w_i, w_j), \tag{58}$$

where w_i is the ith word in the list of ϕ_{wt}, $w \in W$, sorted in descending order. A typical approach is to calculate the top-10 coherence. In addition, we estimated the coherence of top-100 words and the coherence of the topic kernel.

Finally, we define the corresponding measures of kernel size, purity, contrast, and coherence for the topic model by averaging over domain-specific topics $t \in S$.

Text collection In our experiments we used the NIPS dataset, which contains $|D| = 1{,}566$ English articles from the Neural Information Processing Systems conference. The length of the collection in words is $n \approx 2.3 \times 10^6$. The vocabulary size is $|W| \approx 1.3 \times 10^4$. We held out $|D'| = 174$ documents for the testing set. In the preparation step we used BOW toolkit (McCallum 1996) to perform changing to low-case, punctuation elimination, and stop-words removal.

Experimental results In all experiments within this paragraph the number of iterations was set to 40, and the number of topics was set to $|T| = 100$ with $|B| = 10$ background topics.

In Table 1 we compare PLSA (first row), LDA (second row) and multiple regularized topic models. First three columns define a combination of regularizers. Other columns correspond to the quality measures described above.

We use a regularized EM-algorithm with smoothing (23) for LDA model with symmetric Dirichlet prior and usually recommended parameters $\alpha = 0.5,\ \beta = 0.01$.

We use a uniform smoothing for background topics with $\alpha = 0.8,\ \beta = 0.1$.

We use a uniform distribution $\beta_w = \frac{1}{|W|}$ or background distribution $\beta_w = \frac{n_w}{n}$ for sparsing domain-specific topics.

From Table 1 we conclude that the combination of sparsing, smoothing and decorrelation significantly improves all quality measures at once. Sparsing gives up to 98 % zero elements in Φ and 87 % zero elements in Θ. Decorrelation improves purity and coherence. Smoothing helps to transfer common lexis words from domain-specific topics to background topics. A slight loss of the hold-out perplexity is consistent with an observation of Chang et al. (2009) that models which achieve better predictive perplexity often have less interpretable latent spaces.

In experiments we use convergence charts to compare different models and to choose regularization trajectories $\boldsymbol{\tau} = (\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma)$. A convergence chart represents each quality measure of the topic model as a function of the iteration step. These charts give

Table 1 Topic models with various combinations of regularizers: smoothing (Sm), sparsing (Sp) with uniform (u) or background (b) distribution, and decorrelation (Dc)

Sm	Sp	Dc	$\mathscr{P}$	$\mathscr{B}$	$\mathscr{S}_\Phi$	$\mathscr{S}_\Theta$	con	pur	ker	$\mathscr{C}^{\text{ker}}$	$\mathscr{C}^{10}$	$\mathscr{C}^{100}$
−	−	−	1,923	0.00	0.000	0.000	0.43	0.14	100	0.84	0.25	0.17
+	−	−	1,902	0.00	0.000	0.000	0.42	0.12	82	0.93	0.26	0.17
−	u	−	2,114	0.24	0.957	0.867	0.53	0.20	71	0.91	0.25	0.18
−	b	−	2,507	0.51	0.957	0.867	0.46	0.56	151	0.71	0.60	0.58
−	−	+	2,025	0.57	0.561	0.000	0.46	0.38	109	0.82	0.94	0.56
+	u	−	1,961	0.25	0.957	0.867	0.51	0.20	64	**0.97**	0.26	0.18
+	b	−	2,025	0.49	0.957	0.867	0.45	0.52	128	0.77	0.55	0.55
+	−	+	1,985	0.59	0.582	0.000	0.46	0.39	97	0.87	0.93	0.57
+	u	+	2,010	0.73	**0.980**	0.867	**0.56**	0.73	78	**0.94**	0.94	0.62
+	b	+	2,026	**0.80**	**0.979**	0.867	0.52	**0.89**	111	0.81	**0.96**	**0.83**

Quality measures: $\mathscr{P}$—hold-out perplexity, $\mathscr{B}$—background ratio, $\mathscr{S}_\Phi$, $\mathscr{S}_\Theta$—sparsity of matrices Φ, Θ, con—contrast, pur—purity, ker—kernel size, $\mathscr{C}^{\text{ker}}$—kernel coherence, $\mathscr{C}^{10}$, $\mathscr{C}^{100}$—coherence of top 10 and top 100 words. The best values in each column are bold-emphasized

insight into the effects of each regularizer when it is used alone or in combination with others.

Figures 4, 5, and 6 show convergence charts for PLSA and two ARTM regularized models. Quality measures are shown in three charts for each model. The left chart represents a hold-out perplexity $\mathscr{P}$ on the left-hand axis, sparsity $\mathscr{S}_\Phi$, $\mathscr{S}_\Theta$ of matrices Φ, Θ and background ratio $\mathscr{B}$ on the right-hand axis. The middle chart represents kernel size (ker) on the left-hand axis, purity (pur) and contrast (con) on the right-hand axis. The right chart represents the coherence of top10 words $\mathscr{C}^{10}$, top100 words $\mathscr{C}^{100}$, and kernel words $\mathscr{C}^{\text{ker}}$ on the left-hand axis.

Figure 4 shows that PLSA does not sparse matrices Φ, Θ and gives too low topic purity. Also it does not determine background words.

Figure 5 shows the cumulative effect of sparsing domain-specific topics (with background distribution β_w) and smoothing background topics.

Figure 6 shows that decorrelation augments purity and coherence. Also it helps to move common lexis words from the domain-specific topics to the background topics. As a result, the background ratio reaches almost 80 %.

Again, note the important effect of regularization for the ill-posed problem: some of quality measures may change significantly even after the likelihood converges, either with no change or with a slight increase of the perplexity.

Because of the volume limitations we can not show all the convergence charts that we have analyzed in our experiments while choosing a satisfactory regularization trajectory. Below we present only our final recommendations.

It is better to switch on sparsing after the iterative process enters into convergence stage making clear which elements of the matrices Φ, Θ are close to zero. An earlier or a more abrupt sparsing may lead to an increase of perplexity. We enabled sparsing at the 10th iteration and gradually adjusted the regularization coefficient to turn into zeros 8 % of the non-zero elements in each vector θ_d and 10 % in each column ϕ_t per iteration.

Smoothing of the background topics should better start straight from the first iteration, with constant regularization coefficients.

Decorrelation can be activated also from the first iteration, with a maximum regularization coefficient that does not yet significantly increase perplexity. For our collection we chose $\gamma = 2 \times 10^5$.

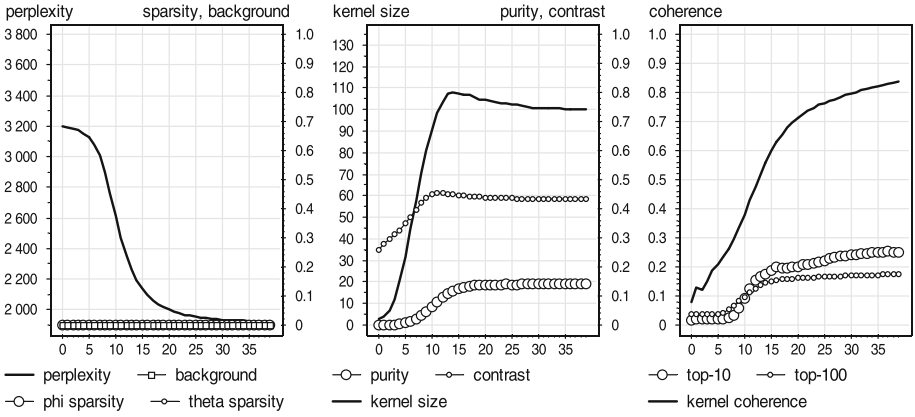

Fig. 4 Convergence charts for PLSA topic model

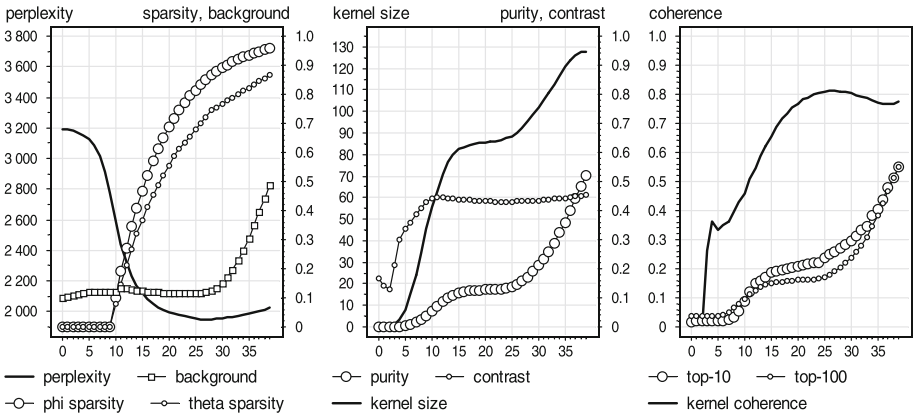

Fig. 5 Convergence charts for ARTM combining sparsing and smoothing

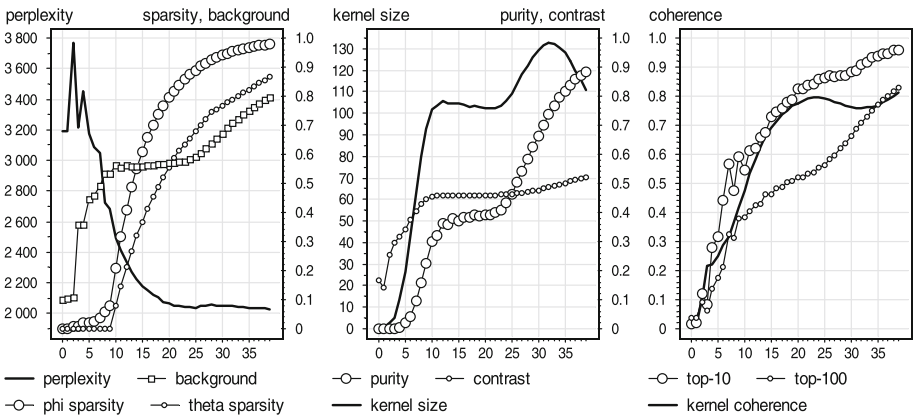

Fig. 6 Convergence charts for ARTM combining sparsing, smoothing, and decorrelation

6 Discussion and conclusions

Learning a topic model from text collection is an ill-posed problem of stochastic matrix factorization. It generally has infinitely many solutions, which is why solutions computed algorithmically are usually unstable and depend on random initialization. Bayesian regularization in the latent Dirichlet allocation does not cope with this problem, indicating that Dirichlet prior is too weak as a regularizer. More problem-oriented regularizers are needed to restrict the set of solutions.

In this paper we propose a semi-probabilistic approach named ARTM—*Additive Regularization of Topic Models*. It is based on the maximization of the weighted sum of the log-likelihood and additional regularization criteria. Learning a topic model is considered as a multi-criteria optimization problem, which then is reduced to a single-criterion problem via scalarization. To solve the optimization problem we use a general regularized EM-algorithm. Compared to the dominant Bayesian approach, ARTM avoids excessive probabilistic assumptions, simplifies the inference of the topic model and allows to use any combination of regularizers.

ARTM provides the theoretical background for developing a library of unified regularizers. With such a library topic models for various applications could be build simply by choosing a suitable combination of regularizers from a pool.

In this paper we introduced a general framework of ARTM under the following constraints, which we intend to remove in further research work.

We confined ourselves to a bag-of-words representation of text collection, and have not considered more sophisticated topic models such as hierarchical, multigram, multilingual, etc. Applying additive regularization to these models will probably require more efforts.

We have worked out only one numerical method—regularized EM-algorithm, suitable for a broad class of regularizers. Alternative optimization techniques as well as their convergence and stability have not yet been considered.

Our review of regularizers is far from being complete. Besides, in our experimental study we have investigated only three of them: sparsing, smoothing, and decorrelation. We argue that this combination improves the interpretability of topics and therefore it is useful for many topic modeling applications. Extensive experiments with combinations of a wider set of regularizers are left beyond the scope of this paper.

Finally, having faced with a problem of regularization trajectory optimization, we confined to a very simple visual technique for monitoring convergence process and comparing topic models empirically.

Acknowledgments The work was supported by the Russian Foundation for Basic Research Grants 14-07-00847, 14-07-00908, 14-07-31176, by Skolkovo Institute of Science and Technology (project 081-R) and by the program of the Department of Mathematical Sciences of Russian Academy of Sciences "Algebraic and combinatoric methods of mathematical cybernetics and information systems of new generation". We thank Alexander Frey and Maria Ryskina for their help and valuable discussions, and Vitaly Glushachenkov for his experimental work on synthetic data.

References

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the international conference on uncertainty in artificial intelligence*, pp. 27–34.

Blei, D., & Lafferty, J. (2007). A correlated topic model of science. *Annals of Applied Statistics*, *1*, 17–35.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural information processing systems (NIPS)*, pp. 288–296.

Chemudugunta, C., Smyth, P., & Steyvers, M. (2007). *Modeling general and specific aspects of documents with a probabilistic topic model* (Vol. 19). Cambridge: MIT Press.

Chien, J. T., & Chang, Y. L. (2013). Bayesian sparse topic model. *Journal of Signal Processessing Systems, 74*(3), 375–389.

Chien, J. T., & Wu, M. S. (2008). Adaptive bayesian latent semantic analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 198–207.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., et al. (2011). TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2412–2421.

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Knowledge discovery through directed probabilistic topic models: A survey. *Frontiers of Computer Science in China*, *4*(2), 280–301.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on machine learning, ICML '07*. ACM, New York, NY, USA, pp. 233–240.

Eisenstein, J., Ahmed, A., & Xing, E. P. (2011). Sparse additive generative models of text. In *ICML '11*, pp. 1041–1048.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*. ACM, New York, NY, USA, pp. 50–57.

Kataria, S., Mitra, P., Caragea, C., & Giles, C. L. (2011). Context sensitive topic models for author influence in document networks. In *Proceedings of the twenty-second international joint conference on artificial intelligence, IJCAI'11*, Vol. 3. AAAI Press, pp. 2274–2280.

Khalifa, O., Corne, D., Chantler, M., & Halley, F. (2013). Multi-objective topic modelling. In *7th International conference evolutionary multi-criterion optimization (EMO 2013)*. Springer LNCS, pp. 51–65.

Larsson, M. O., & Ugander, J. (2011). A concave regularization technique for sparse mixture models. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, pp. 1890–1898.

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, *14*(2), 178–203.

Mann, G. S., & McCallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proceedings of the 24th international conference on machine learning, ICML '07*. ACM, New York, NY, USA, pp. 593–600.

Masada, T., Kiyasu, S., & Miyahara, S. (2008). Comparing LDA with pLSI as a dimensionality reduction method in document clustering. In *Proceedings of the 3rd international conference on large-scale knowledge resources: construction and application, LKR'08*. Springer, Berlin, pp. 13–26.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/mccallum/bow

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing, association for computational linguistics, EMNLP '11*, Stroudsburg, PA, USA, pp. 262–272.

Newman, D., Karimi, S., & Cavedon, L. (2009). External evaluation of topic models. In *Australasian document computing symposium*, pp. 11–18.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010a). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, HLT '10*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 100–108.

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010b). Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on digital libraries, JCDL '10*. ACM, New York, NY, USA, pp. 215–224.

Newman, D., Bonilla, E. V., & Buntine, W. L. (2011). Improving topic coherence with regularized topic models. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems*, Vol. 24, pp. 496–504.

Potapenko, A. A., & Vorontsov, K. V. (2013). Robust PLSA performs better than LDA. In *35th European conference on information retrieval*, ECIR-2013, Moscow, Russia, 24–27 March 2013, *Lecture notes in computer science (LNCS)*. Springer, Germany, pp. 784–787.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence, UAI '04*. AUAI Press, Arlington, VA, USA, pp. 487–494.

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning*, *88*(1–2), 157–208.

Shashanka, M., Raj, B., & Smaragdis, P. (2008). Sparse overcomplete latent variable decomposition of counts data. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems, NIPS-2007* (pp. 1313–1320). Cambridge, MA: MIT Press.

Si, L., & Jin, R. (2005). Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis. In T. B. Ho, D. W. L. Cheung, & H. Liu (Eds.), *Proceedings of the ninth Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Lecture notes in computer science*, Vol. 3518. Springer, Berlin, pp. 622–631.

Steyvers, M., & Griffiths, T. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5228–5235.

Tan, Y., & Ou, Z. (2010). Topic-weak-correlated latent Dirichlet allocation. In *7th International symposium Chinese spoken language processing (ISCSLP)*, pp. 224–228.

Teh, Y. W., Newman, D., & Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *NIPS*, pp. 1353–1360.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solution of ill-posed problems*. Washington, DC: W. H. Winston.

Varadarajan, J., Emonet, R., & Odobez, J. M. (2010). A sparsity constraint for topic models—Application to temporal activity mining. In *NIPS-2010 workshop on practical applications of sparse modeling: Open issues and new directions*.

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22: 23rd Annual Conference on Neural Information Processing Systems*. Vancouver, BC, Canada, pp. 1973–1981.

Wang, C., & Blei, D. M. (2009). Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *NIPS*. Curran Associates, Inc., pp. 1982–1989.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, New York, NY, USA, pp. 448–456.

Wang, Q., Xu, J., Li, H., & Craswell, N. (2011). Regularized latent semantic indexing. In *SIGIR*, pp. 685–694.

Wu, Y., Ding, Y., Wang, X., & Xu, J. (2010). A comparative study of topic models for topic clustering of Chinese web news. In *2010 3rd IEEE international conference on computer science and information technology (ICCSIT)*, Vol. 5, pp. 236–240.

Zhou, S., Li, K., & Liu, Y. (2009). Text categorization based on topic model. *International Journal of Computational Intelligence Systems*, *2*(4), 398–409.