Additive Regularization of Topic Models for Topic Selection and Sparse Factorization

Konstantin Vorontsov[1], Anna Potapenko[2], and Alexander Plavin[3]

[1] Moscow Institute of Physics and Technology,
Dorodnicyn Computing Centre of RAS,
National Research University Higher School of Economics
voron@forecsys.ru
[2] National Research University Higher School of Economics
anya_potapenko@mail.ru
[3] Moscow Institute of Physics and Technology
alexander@plav.in

Abstract. Probabilistic topic modeling of text collections is a powerful tool for statistical text analysis. Determining the optimal number of topics remains a challenging problem in topic modeling. We propose a simple entropy regularization for topic selection in terms of *Additive Regularization of Topic Models* (ARTM), a multicriteria approach for combining regularizers. The entropy regularization gradually eliminates insignificant and linearly dependent topics. This process converges to the correct value on semi-real data. On real text collections it can be combined with sparsing, smoothing and decorrelation regularizers to produce a sequence of models with different numbers of well interpretable topics.

Keywords: probabilistic topic modeling, regularization, Probabilistic Latent Sematic Analysis, topic selection, EM-algorithm

1 Introduction

Topic modeling is a rapidly developing branch of statistical text analysis (Blei, 2012). Topic model reveals a hidden thematic structure of the text collection and finds a highly compressed representation of each document by a set of its topics. From the statistical point of view, a probabilistic topic model defines each topic by a multinomial distribution over words, and then describes each document with a multinomial distribution over topics. Such models appear to be highly useful for many applications including information retrieval, classification, categorization, summarization and segmentation of texts. More ideas and applications are outlined in the survey (Daud et al, 2010).

Determining an appropriate number of topics for a given collection is an important problem in probabilistic topic modeling. Choosing too few topics results in too general topics, while choosing too many ones leads to insignificant and highly similar topics. *Hierarchical Dirichlet Process*, HDP (Teh et al, 2006; Blei

et al, 2010) is the most popular Bayesian approach for number of topics optimization. Nevertheless, HDP sometimes gives very unstable number of topics and requires a complicated inference if combined with other models.

To address the above problems we use a non-Bayesian semi-probabilistic approach — *Additive Regularization of Topic Models*, ARTM (Vorontsov, 2014; Vorontsov and Potapenko, 2014a). 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. In order to choose a better solution, we maximize the log-likelihood with a weighted sum of regularization penalty terms. These regularizers formalize additional requirements for a topic model. Unlike Bayesian approach, ARTM avoids excessive probabilistic assumptions and simplifies the inference of multi-objective topic models.

The aim of the paper is to develop topic selection technique for ARTM based on entropy regularization and to study its combinations with other useful regularizers such as sparsing, smoothing and decorrelation.

The rest of the paper is organized as follows. In section 2 we introduce a general ARTM framework, the regularized EM-algorithm, and a set of regularizers including the entropy regularizer for topic selection. In section 3 we use semi-real dataset with known number of topics to show that the entropy regularizer converges to the correct number of topics, gives a more stable result than HDP, and gradually removes linearly dependent topics. In section 4 the experiments on real dataset give an insight that optimization of the number of topics is in its turn an ill-posed problem and has many solutions. We propose additional criteria to choose the best of them. In section 5 we discuss advantages and limitations of ARTM with topic selection regularization.

2 Additive Regularization of Topic Models

Let D denote a set (collection) of texts and W denote a set (vocabulary) of all terms that appear in these texts. A term can be a single word or a keyphrase. Each document $d \in D$ is a sequence of n_d terms $(w_1, \ldots, w_{n_d})$ from W. Denote n_{dw} the number of times the term w appears in the document d.

Assume that each term occurrence in each document refers to some latent topic from a finite set of topics T. Then text collection is considered as 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 $W \times D \times T$. Terms w and documents d are observable variables, while topics t are latent variables. Following the "bag of words" model, we represent each document as a subset of terms $d \subset W$.

A probabilistic topic model describes how terms of a document are generated from a mixture of given distributions $\phi_{wt} = p(w \mid t)$ and $\theta_{td} = p(t \mid d)$:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t \in T} \phi_{wt} \theta_{td}. \tag{1}$$

Learning a topic model is an inverse problem to find distributions ϕ_{wt} and θ_{td} given a collection D. This problem is equivalent to finding an approximate representation of frequency matrix $F = \left(\frac{n_{dw}}{n_d}\right)_{W \times D}$ with a product $F \approx \Phi\Theta$ of two

unknown matrices — the matrix $\Phi = (\phi_{wt})_{W\times T}$ of *term probabilities for the topics* and the matrix $\Theta = (\theta_{td})_{T\times D}$ of *topic probabilities for the documents*. Matrices F, Φ and Θ are *stochastic*, that is, their columns are non-negative, normalized, and represent discrete distributions. Usually $|T| \ll |D|$ and $|T| \ll |W|$.

In *Probabilistic Latent Semantic Analysis*, PLSA (Hofmann, 1999) a topic model (1) is learned by log-likelihood maximization with linear constrains:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{2}$$

$$\sum_{w\in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \qquad \sum_{t\in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \tag{3}$$

The product $\Phi\Theta$ is defined up to a linear transformation $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, where matrices $\Phi' = \Phi S$ and $\Theta' = S^{-1}\Theta$ are also stochastic. Therefore, in a general case the maximization problem (2) has an infinite set of solutions.

In *Additive Regularization of Topic Models*, ARTM (Vorontsov, 2014) a topic model (1) is learned by maximization of a linear combination of the log-likelihood (2) and r regularization penalty terms $R_i(\Phi,\Theta)$, $i = 1,\ldots,r$ with nonnegative *regularization coefficients* τ_i:

$$R(\Phi,\Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi,\Theta), \qquad L(\Phi,\Theta) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}. \tag{4}$$

The Karush–Kuhn–Tucker conditions for (4), (3) give (under some technical restrictions) the necessary conditions for the local maximum in a form of the system of equations (Vorontsov and Potapenko, 2014a):

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}; \tag{5}$$

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+; \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{6}$$

$$\theta_{td} \propto \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+; \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}; \tag{7}$$

where $(z)_+ = \max\{z, 0\}$. Auxiliary variables p_{tdw} are interpreted as conditional probabilities of topics for each word in each document, $p_{tdw} = p(t \mid d, w)$.

The system of equations (5)–(7) can be solved by various numerical methods. Particularly, the simple-iteration method is equivalent to the EM algorithm, which is typically used in practice. The pseudocode of Algorithm 2.1 shows its rational implementation, in which E-step (5) is incorporated into M-step (6)–(7), thus avoiding storage of 3D-array p_{tdw}.

The strength of ARTM is that each additive regularization term results in a simple additive modification of the M-step. Many models previously developed within Bayesian framework can be easier reinterpreted, inferred and combined using ARTM framework (Vorontsov and Potapenko, 2014a,b).

Algorithm 2.1: The regularized EM-algorithm for ARTM.

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

initialize vectors ϕ_t, θ_d randomly;
repeat
  $n_{wt} := 0$, $n_{td} := 0$ for all $d \in D,\ w \in W,\ t \in T$;
  for all $d \in D,\ w \in d$
    $p(w \mid d) := \sum_{t \in T} \phi_{wt}\theta_{td}$;
    increase n_{wt} and n_{td} by $n_{dw}\phi_{wt}\theta_{td}/p(w \mid d)$ for all $t \in T$;
  $\phi_{wt} \propto \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+$ for all $w \in W,\ t \in T$;
  $\theta_{td} \propto \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+$ for all $t \in T$, $d \in D$;
until Φ *and* Θ *converge*;

To find a reasonable number of topics we propose to start from a wittingly large number and gradually eliminate insignificant or excessive topics from the model. To do this we perform the entropy-based sparsing of distribution $p(t) = \sum_d p(d)\theta_{td}$ over topics by maximizing KL-divergence between $p(t)$ and the uniform distribution over topics (Vorontsov and Potapenko, 2014b):

$$R(\Theta) = \frac{n}{|T|}\sum_{t \in T} \ln \sum_{d \in D} p(d)\theta_{td} \ \to\ \max.$$

Substitution of this regularizer into the M-step equation (7) gives

$$\theta_{td} \propto \left(n_{td} - \tau\frac{n}{|T|}\frac{n_d}{n_t}\theta_{td}\right)_+.$$

Replacing θ_{td} in the right-hand side by its unbiased estimate $\frac{n_{td}}{n_d}$ gives an interpretation of the regularized M-step as a row sparser for the matrix Θ:

$$\theta_{td} \propto n_{td}\left(1 - \tau\frac{n}{|T|n_t}\right)_+.$$

If n_t counter in the denominator is small, then all elements of a row will be set to zero, and the corresponding topic t will be eliminated from the model. Values τ are normally in $[0, 1]$ due to the normalizing factor $\frac{n}{|T|}$.

Our aim is to understand how the entropy-based topic sparsing works and to study its behavior in combinations with other regularizers. We use a set of three regularizers — sparsing, smoothing and decorrelation proposed in (Vorontsov and Potapenko, 2014a) to divide topics into two types, $T = S \sqcup B$: domain-specific topics S and background topics B.

Domain-specific topics $t \in S$ contain terms of domain areas. They are supposed to be sparse and weakly correlated, because a document is usually related to a small number of topics, and a topic usually consists of a small number of domain-specific terms. Sparsing regularization is based on KL-divergence maximization between distributions ϕ_{wt}, θ_{td} and corresponding uniform distributions.

Decorrelation is based on covariance minimization between all topic pairs and helps to exclude common lexis from domain-specific topics (Tan and Ou, 2010).

Background topics $t \in B$ contain common lexis words. They are smoothed and appear in many documents. Smoothing regularization minimizes KL-divergence between distributions ϕ_{wt}, θ_{td} and corresponding uniform distributions. Smoothing regularization is equivalent to a maximum a posteriori estimation for LDA, *Latent Dirichlet Allocation* topic model (Blei et al, 2003).

The combination of all mentioned regularizers leads to the M-step formulas:

$$\phi_{wt} \propto \Big(n_{wt} - \beta_0 \underbrace{\beta_w[t \in S]}_{\substack{\text{sparsing}\\ \text{specific}\\ \text{topic}}} + \beta_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{smoothing}\\ \text{background}\\ \text{topic}}} - \underbrace{\gamma\,[t \in S]\,\phi_{wt} \sum_{s \in S \setminus t} \phi_{ws}}_{\text{topic decorrelation}}\Big)_+; \tag{8}$$

$$\theta_{td} \propto \Big(n_{td} - \alpha_0 \underbrace{\alpha_t[t \in S]}_{\substack{\text{sparsing}\\ \text{specific}\\ \text{topic}}} + \alpha_1 \underbrace{\alpha_t[t \in B]}_{\substack{\text{smoothing}\\ \text{background}\\ \text{topic}}} - \underbrace{\tau\,[t \in S]\frac{n}{|T|}\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{topic}\\ \text{selection}}}\Big)_+; \tag{9}$$

where regularization coefficients α_0, α_1, β_0, β_1, γ, τ are selected experimentally, distributions α_t and β_w are uniform.

3 Number of Topics Determination

In our experiments we use NIPS dataset, which contains $|D| = 1740$ English articles from the Neural Information Processing Systems conference for 12 years. We use the version, preprocessed by A. McCallum in BOW toolkit (McCallum, 1996), where changing to low-case, punctuation elimination, and stop-words removal were performed. The length of the collection in words is $n \approx 2.3 \cdot 10^6$ and the vocabulary size is $|W| \approx 1.3 \cdot 10^4$.

In order to assess how well our approach determines the number of topics, we generate semi-real (synthetic but realistic) datasets with the known number of topics. First, we run 500 EM iterations for PLSA model with T_0 topics on NIPS dataset and generate synthetic dataset $\Pi_0 = (n^0_{dw})$ from Φ, Θ matrices of the solution: $n^0_{dw} = n_d \sum_{t \in T} \phi_{wt}\theta_{td}$. Second, we construct a parametric family of semi-real datasets $\Pi_\alpha = (n^\alpha_{dw})$ as a mixture $n^\alpha_{dw} = \alpha n_{dw} + (1-\alpha)n^0_{dw}$, where $\Pi_1 = (n_{dw})$ is the term counters matrix of the real NIPS dataset.

From synthetic to real dataset. Fig. 1 shows the dependence of revealed number of topics on the regularization coefficient τ for two families of semi-real datasets, obtained with $T_0 = 50$ and $T_0 = 25$ topics. For synthetic datasets ARTM reliably finds the true number of topics for all τ in a wide range. Note, that this range does not depend much on the number of topics T_0, chosen for datasets generation. Therefore, we conclude that an approximate value of regularization coefficient $\tau = 0.25$ from the middle of the range is recommended for determining number of topics via our approach.

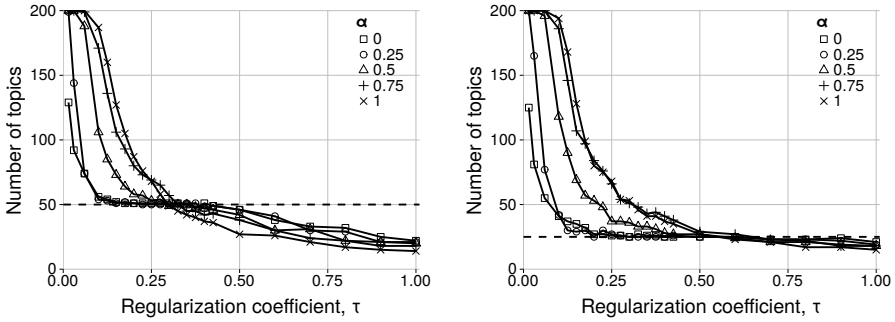


Fig. 1. ARTM for semi-real datasets with $T_0 = 50$ (left) and $T_0 = 25$ (right).

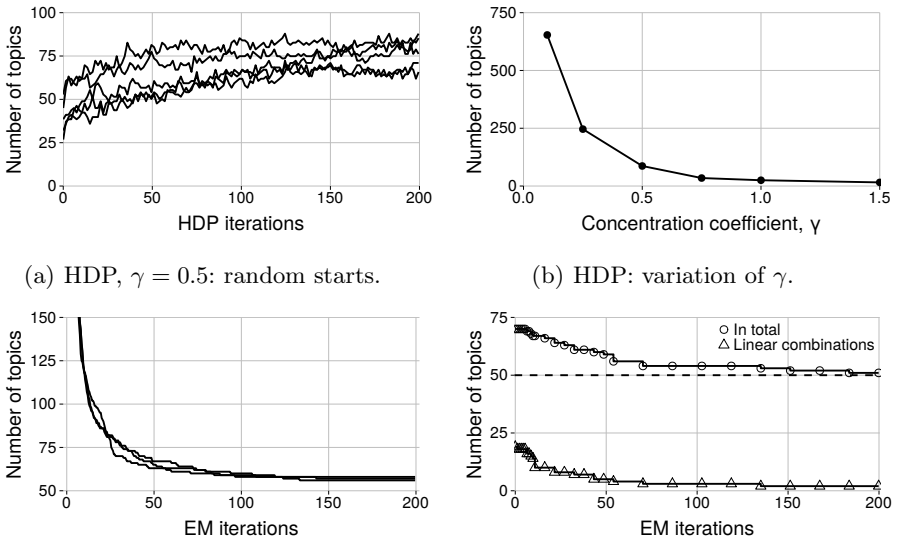
However as the data changes from synthetic Π_0 to real Π_1, the horizontal part of the curve diminishes, and for NIPS dataset there is no evidence for the "best" number of topics. This corresponds to the intuition that real text collections do not expose the "true number of topics", but can be reasonably described by models with different number of topics.

Comparison of ARTM and HDP models. In our experiments we use the implementation[1] of HDP by C. Wang and D. Blei. Fig. 2(b) demonstrates that the revealed number of topics depends on the parameter of the model not only for ARTM approach (Fig. 1, $\alpha = 1$ case), but for HDP as well. Varying the concentration coefficient γ of Dirichlet process, we can get any number of topics.

Fig. 2(a) presents a bunch of curves, obtained for several random starts of HDP with default $\gamma = 0.5$. Here we observe the instability of the method in two ways. Firstly, there are incessant fluctuations of number of topics from iteration to iteration. Secondly, the results for several random starts of the algorithm significantly differ. Comparing Fig. 2(a) and Fig. 2(c) we conclude that our approach is much more stable in both ways. The numbers of topics, determined by two approaches with recommended values of parameters, are similar.

Elimination of linearly dependent topics. One more important question is which topics are selected for exclusion from the model. To work it out, we extend the synthetic dataset Π_0 to model linear dependencies between the topics. 50 topics obtained by PLSA are enriched by 20 convex combinations of some of them; and new vector columns are added to Φ matrix. The corresponding rows in Θ matrix are filled with random values drawn from a bag of elements of original Θ, in order to make values in the new rows similarly distributed. These matrices are then used as synthetic dataset for regularized EM-algorithm with topic selection to check whether original or combined topics remain. Fig. 2(d) demonstrates that the topic selection regularizer eliminates excessive linear combinations, while more sparse and diverse topics of the original model remain.

[1] `http://www.cs.princeton.edu/∼chongw/resource.html`.



(a) HDP, $\gamma = 0.5$: random starts.

(b) HDP: variation of γ.

(c) ARTM, $\tau = 0.25$: random starts.

(d) ARTM: topic selection.

Fig. 2. ARTM and HDP models for determining number of topics.

4 Topic Selection in a Sparse Decorrelated Model

The aim of the experiments in this section is to show that the proposed topic selection regularizer works well in combination with other regularizers. The topic model quality is evaluated by multiple criteria.

The *hold-out perplexity* $\mathscr{P} = \exp\bigl(-\frac{1}{n}L(\Phi,\Theta)\bigr)$ is the exponential average of the likelihood on a test set of documents; the lower, the better.

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

The *background ratio* $\mathscr{B} = \frac{1}{n}\sum_{d\in D}\sum_{w\in d}\sum_{t\in B} n_{dw}p(t\,|\,d,w)$ is a ratio of background terms over the collection. It takes values from 0 to 1. If $\mathscr{B} \to 0$ then the model doesn't distinguishes common lexis from domain-specific terms. If $\mathscr{B} \to 1$ then the model is degenerated, possibly due to excessive sparsing.

The *lexical kernel* W_t of a topic t is a set of terms that distinguish the topic t from the others: $W_t = \{w\colon p(t\,|\,w) > \delta\}$. In our experiments $\delta = 0.25$. We use the notion of lexical kernel to define two characteristics of topic interpetability.

The *purity* $\sum_{w\in W_t} p(w\,|\,t)$ shows the cumulative ratio of kernel in the topic.

The *contrast* $\frac{1}{|W_t|}\sum_{w\in W_t} p(t\,|\,w)$ shows the diversity of the topic.

The *coherence of a topic* $\mathscr{C}_t^k = \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=i}^{k}\mathrm{PMI}(w_i,w_j)$ is defined as the average pointwise mutual information over word pairs, where w_i is the i-th word in the list of k most probable words in the topic. Coherence is commonly used as the interpretability measure of the topic model (Newman et al, 2010). We estimate the coherence for top-10, top-100, and besides, for lexical kernels.

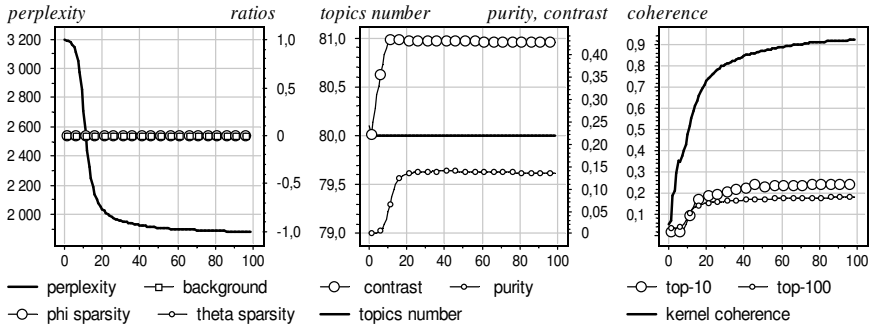


Fig. 3. Baseline: LDA topic model.

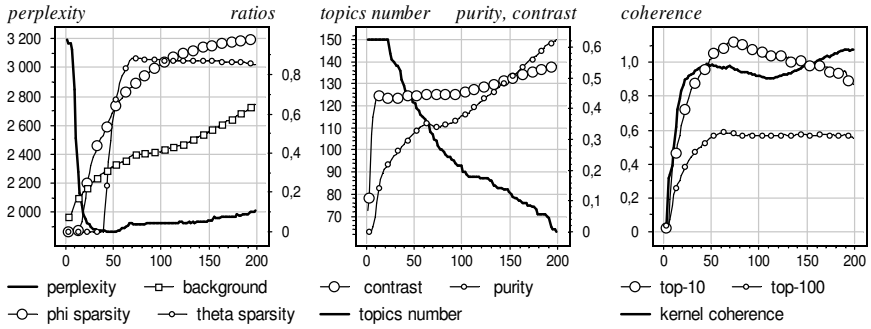


Fig. 4. Combination of sparsing, decorrelation, and topic selection.

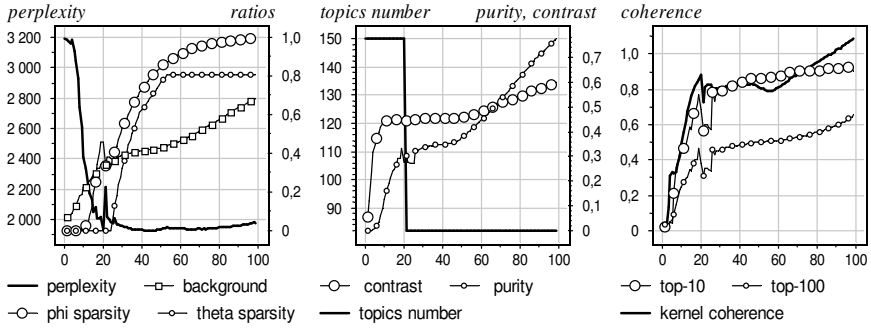


Fig. 5. Sequential phases of regularization.

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

Further we represent each quality measure of the topic model as a function of the iteration step and use several charts for better visibility. Fig. 3 provides such charts for a standard LDA model, while Fig. 5 and Fig. 4 present regularized models with domain-specific and background topics. We use constant parameters for smoothing background topics $|S| = 10,\ \alpha_t = 0.8,\ \beta_w = 0.1$.

The model depicted in Fig. 4 is an example of simultaneous sparsing, decorrelating and topic selection. Decorrelation coefficient grows linearly during the first 60 iterations up to the highest value $\gamma = 200000$ that does not deteriorate the model. Topic selection with $\tau = 0.3$ is turned on later, after the 15-th iteration. Topic selection and decorrelation are used at alternating iterations because their effects may conflict; in charts we depict the quality measures after decorrelating iterations. To get rid of insignificant words in topics and to prepare insignificant topics for further elimination, sparsing is turned on staring from the 40-th iteration. Its coefficients α_t, β_w gradually increase to zeroize 2% of Θ elements and 9% of Φ elements each iteration. As a result, we get a sequence of models with decreasing number of sparse interpretable domain-specific topics: their purity, contrast and coherence are noticeably better then those of LDA topics.

Another regularization strategy is presented in Fig. 5. In contrast with the previous one, it has several sequential phases for work of different regularizers. Firstly, decorrelation makes topics as different as possible. Secondly, topic selection eliminates excessive topics and remains 80 topics of 150. Note, that in spite of small $\tau = 0.1$, many topics are excluded at once due to the side effect of the first phase. The remained topics are significant, and none of them manage to be excluded later on. The final phase performs both sparsing and decorrelating of the remained topics to successfully improve their interpretability.

It is curious that the number of topics 80, determined by this strategy, corresponds to the results of the previous strategy quite well. In Fig. 4 we observe two regions of perplexity deterioration. The first one concerns Θ sparsing; after that the perplexity remains stable for a long period till the 150-th iteration, when the number of topics becomes less than 80. This moment indicates that all the remained topics are needed and should not be further eliminated.

5 Conclusions

Learning a topic model from text collection is an ill-posed problem of stochastic matrix factorization. Determining the number of topics is an ill-posed problem too. In this work we develop a regularization approach to topic selection in terms of non-Bayesian ARTM framework. Starting with excessively high number of topics we gradually make them more and more sparse and decorrelated, and eliminate unnecessary topics by means of entropy regularization. This approach gives more stable results than HDP and during one learning process generates a sequence of models with quality measures trade-off. The main limitation, which should be removed in future work, is that regularization coefficients are not optimized automatically, and we have to choose the regularization strategy manually.

Acknowledgements. The work was supported by the Russian Foundation for Basic Research grants 14-07-00847, 14-07-00908, 14-07-31176, Skolkovo Institute of Science and Technology (project 081-R), and by the program of the Department of Mathematical Sciences of RAS “Algebraic and combinatoric methods of mathematical cybernetics and information systems of new generation”.

Bibliography

Blei DM (2012) Probabilistic topic models. Communications of the ACM 55(4):77–84

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022

Blei DM, Griffiths TL, Jordan MI (2010) The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. J ACM 57(2):7:1–7:30

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

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

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

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

Tan Y, Ou Z (2010) Topic-weak-correlated latent dirichlet allocation. In: 7th International Symposium Chinese Spoken Language Processing (ISCSLP), pp 224–228

Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476):1566–1581

Vorontsov KV (2014) Additive regularization for topic models of text collections. Doklady Mathematics 89(3):301–304

Vorontsov KV, Potapenko AA (2014a) Additive regularization of topic models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization

Vorontsov KV, Potapenko AA (2014b) Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: AIST'2014, Analysis of Images, Social networks and Texts, Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), vol 436, pp 29–46