Analyzing the Influence of Hyper-parameters and Regularizers of Topic Modelling in Terms of Renyi Entropy

Anonymous Author(s)

Abstract

Topic modelling is a popular approach for clustering text documents. A variety of different types of regularization is implemented in topic modelling. In this paper we propose a novel approach for analyzing the influence of different regularization types on results of topic modelling. Based on Renyi entropy, this approach is inspired by the concepts from statistical physics, where an inferred topical structure of a collection can be considered an information statistical system residing in a non-equilibrium state. By testing our approach on three models - Probabilistic Latent Semantic Analysis (pLSA), Additive Regularization of Topic Models (BigARTM) and Latent Dirichlet Allocation (LDA) with Gibbs sampling - we, first, show that the minimum of Renyi entropy coincides with the "true" number of topics, as determined in two labelled collections. Simultaneously we find that Hierarchical Dirichlet Process (HDP) model as a well-known approach for topic number optimization fails to detect such optimum. Next, we demonstrate that large values of the regularization coefficient in BigARTM significantly shift the minimum of entropy from the topic number optimum, which effect is not observed for hyper-parameters in LDA. We conclude that regularization may introduce unpredictable distortions into topic models that need further research.

1 Introduction

Topic modelling (TM) is a popular statistical approach for discovering latent topics in a collection of documents, where each topic is a distribution over the vocabulary. Most of the existing TM models are based on different types of regularization and, hence, are controlled by regularization penalty terms (e.g. BigARTM) or hyper-parameters (e.g. LDA). It has been proved that the choice of these parameters has a large impact on the modelling [George and Doss, 2017]. Estimating their impact on the result of TM and searching for their optimal values are not trivial. Moreover, existing metrics in the field of TM are time-consuming and usually based on monotonous functions such as exponent or logarithm that do not effectively assist the determination of the set of parameters. In this paper we propose an effective method, based on Renyi entropy [Renyi, 1970], for analyzing the influence of regularization on the outcome of TM. Our approach also allows us to estimate optimal values of topic model parameters including the number of topics and regularization parameters. We compare the results of our approach with log-likelihood metric and find that our method is faster and, in addition, allows to estimate the optimal number of topics while log-likelihood does not.

2 Background

TM is a family of mathematical algorithms based on the following assertions [Hofmann, 1999]:

1. Let $\hat{D}$ be a collection of textual documents with D documents, $\hat{W}$ be a set of all unique terms (vocabulary) with W elements. 2. It is assumed that there exists a finite number T of topics, and each entry of a word w in document d is associated with a certain topic $t \in \hat{T}$ ($\hat{T}$ is a set of topics). 3. A collection of documents is considered stochastic independent sample of triples (w_i, d_i, t_i), $i = 1, ..., n$, from a discrete distribution $p(w, d, t)$ on a finite probability space $\hat{W} \times \hat{D} \times \hat{T}$. Words and documents are observable variables, topics are latent (hidden) variables. 4. It is assumed that the order of words in the set of documents is not important for TM ('bag-of-words' model). Similarly, the order of documents in a collection is also insignificant.

In TM, the probability $p(w|d)$ of a term w to occur in a document d can be expressed as follows:

$p(w|d) = \sum_{t \in \hat{T}} p(w|t)p(t|d) = \sum_{t \in \hat{T}} \phi_{wt}\theta_{td}$, where $p(w|t) = \phi_{wt}$ is the probability of a word w to occur under a topic t, $p(t, d) = \theta_{td}$ is the probability of a topic t in a document d. Probabilities ϕ_{wt} form a matrix of distribution of words by topics $\Phi = (\phi_{wt})_{w \in \hat{W}, t \in \hat{T}}$ and probabilities θ_{td} form a matrix of distribution of topics by documents $\Theta = (\theta_{td})_{t \in \hat{T}, d \in \hat{D}}$. Nowadays, several types of models exist in the field of TM, and can be classified into three categories: 1. Topic models based on maximum likelihood principle [Blei *et al.*, 2003]. Here matrices Φ and Θ are searched by Expectation-Maximization (E-M) algorithm. 2. Topic models based on Markov chains (Gibbs sampling model) [Asuncion *et al.*, 2009], where ϕ_{wt} and θ_{td} are searched by calculating expectation through Monte-Carlo method. De-

spite different mathematical approaches of these two types of models, both of them produce similar topic solutions [Asuncion *et al.*, 2009]. 3. Hierarchical Dirichlet Process (HDP) is an alternative model which is considered in literature as a non-parametric [Wang *et al.*, 2011]. However, it admits a set of predefined parameters (e.g. truncation level) which influence the results of modelling. In our paper, HDP model was used for estimating the "optimal" number of topics. Let us note that in the process of TM for models based on both E-M algorithm or Gibbs sampling (GS) algorithm, a transition occurs to a highly non-equilibrium state (or non-uniform state, in other words). The flat distribution is chosen as the initial distribution of words by topics and documents by topics for LDA (GS), while for pLSA (E-M) and LDA (E-M) the initial distribution is defined by random number generator. For both types of algorithms the initial distribution corresponds to maximum entropy. In the process of TM, for all types of algorithms and initial distributions, redistribution of words and documents by topics proceeds so that a significant portion of words (about 95% of all unique words) acquires probabilities close to zero and only about 3-5 % of words receive relatively high probabilities [Koltcov *et al.*, 2014]. Numerical experiments demonstrate that the number of words with high probabilities depends on the number of topics and values of hyperparameters that allows us to construct a theoretical approach for analyzing such dependency using a perspective of statistical physics. Now let us discuss log-likelihood and perplexity known as standard metrics in TM. The log-likelihood of a set of documents can be directly expressed as a function of the TM parameters. For example, for LDA model one obtains the following expression [Wallach *et al.*, 2009], [Heinrich, 2004]: $\ln(P(\hat{D}|\Phi,\alpha) = \sum_{d=1}^{D}\sum_{w=1}^{W} n_{dw}\ln(\sum_{t=1}^{T}\phi_{wt}\theta_{td})$, where n_{dw} is frequency of word w in document d, α is hyperparameter for topic distribution in documents. A better model will yield higher probabilities of documents, on average [Wallach *et al.*, 2009]. Another measure which is closely related to likelihood is perplexity which is defined for LDA model as: Perplexity $= \exp(-\ln(P(\hat{D}|\Phi,\alpha)/\sum_{d=1}^{D} n_d)$, where n_d is the number of words in document d. Perplexity behaves as a monotone decreasing function [Asuncion *et al.*, 2009] and is dependent on the size of vocabulary [De Waal and Barnard, 2008] which makes it unsuitable for comparison across datasets. Therefore, in our work we use only likelihood.

3 Entropy approach for analyzing topic models

Our entropy approach is based on the following assertions [Koltcov, 2018]: 1. A collection of documents is considered a mesoscopic information system consisting of millions of elements (words and documents). Correspondingly, behavior of such a system can be studied by application of models from statistical physics. 2. A topic is a state (an analogue of spin direction), which each word and document in the collection can take. Here, a word and a document can belong to different topics with different probabilities. For example, a word w resides in state t with probability p_{wt}, so, elements of matrix Φ are probabilities of micro-states in terms of physics. 3. Such information system is open and exchanges energy with the environment via changing the temperature. Here, the temperature of information system is the number of topics which is a parameter and should be selected by searching for a minimum non-extensive entropy of the system. 4. To measure the degree to which a given system is non-equilibrium one can use the difference of free energies $\Lambda_F = F(T) - F_0$, where F_0 is the free energy of the initial state (chaos), $F(T)$ is the free energy calculated after TM for a given number of topics. 5. The values of hyper-parameters or regularization coefficients and the number of topics are varied parameters, which influence the value of the difference of free energies. 6. The optimal number of topics and the set of optimal hyperparameters of topic model corresponds to the situation when information maximum is reached (i.e. free energy minimum).

The sum of probabilities of words in a topic model equals to the number of topics, i.e., $T = \sum_{t=1}^{T}\sum_{w=1}^{W} p_{wt}$, where $p_{wt} \in [0,1]$ for all $w = 1,..,W;\ t = 1,...,T$. In the framework of statistical physics it is common to investigate distribution of statistical systems by energy levels, where energy is expressed in terms of probability. In accordance with this approach, we divide the range of probabilities $[0,1]$ by a fixed number of intervals and determine energy levels corresponding to these intervals, and then seek the number of micro-states belonging to each energy level (note that these numbers depend on the number of topics and values of hyperparameters of a topic model). Division into intervals is convenient from a computational point of view. If the lengths of such intervals tend to zero, the distribution of words by intervals will tend to probability density function. However, for simplification we will consider a two-level system, where the first level corresponds to words with small probabilities close to zero and the second level corresponds to words with high probabilities.

On this basis, we introduce 'density-of-states function' for words with high probabilities as follows: $\rho = N/(WT)$, where N is the number of micro-states with high probabilities. By high probability we mean probability satisfying: $p > 1/W$. The choice of such level is informed with the fact that values $1/W$ are the initial values of matrix Φ for a topic model. Hence, during the process of TM, probabilities of words redistribute with respect to this threshold level. Based on the concepts of statistical physics, the level of micro-states with high probabilities can be characterized by the amount of energy expressed in terms of combination of probabilities of micro-states residing in the given interval (in our case, above the threshold level $1/W$):

$$E = -\ln(\tilde{P}) = -\ln(\sum_{wt} p_{wt}\cdot\Omega(p_{wt} - 1/W)), \qquad (1)$$

where the step function Ω is defined by $\Omega(p_{wt} - 1/W) = 1$ if $p_{wt} \geq 1/W$, $\Omega(p_{wt} - 1/W) = 0$ if $p_{wt} < 1/W$. So, in equation (1) we sum only probabilities that are greater than $1/W$. The energy level is characterized by two parameters: 1. Sum of probabilities of micro-states, that lie in the corresponding interval, $\tilde{P}$; 2. The number of micro-states, N, whose probabilities lie in this interval. For a two-leveled system the main contribution to the entropy and energy of the whole system is given by the states with high probabilities,

that is mainly by the upper level. Respectively, the free energy of the whole system is almost entirely determined by the entropy and the energy of one level. Free energy of a statistical system can be expressed through Gibbs-Shannon entropy and through the internal energy in the following way [Tsallis, 2009]: $F = E - TS = E - S/q$, where $q = 1/T$. Entropy of an information statistical system can be expressed through the number of micro-states belonging to the same level [Tkačik *et al.*, 2015] : $S = \ln(N)$. It follows that free energy of a topic model is expressed through $\tilde{P}$ and ρ in the following way:

$$\begin{aligned}\Lambda_F =& F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0)T = \\ =& -\ln(\tilde{P}/T) - T\ln(\rho),\end{aligned} \tag{2}$$

where E_0, S_0 are the energy and the entropy of the initial state of the system. Hence, the degree to which a given system is non-equilibrium can be defined as the difference between the two free energies and expressed in terms of experimentally determined values ρ and $\tilde{P}$. These values are calculated for each topic model under variation of parameter T and hyper-parameters.

On the other hand, free energy and Renyi entropy can be expressed in terms of partition function (statistical sum). The latter, in turn, can by expressed in terms of ρ and $\tilde{P}$ [Mora and Walczak, 2016]: $Z_q = e^{-q\Lambda_F} = e^{-qE+S} = \rho(q\tilde{P})^q$, where $q = 1/T$. This relation allows us to express Renyi entropy in terms of free energy and experimentally determined values $\tilde{P}$ and ρ:

$$S_q^R = \frac{\ln(Z_q)}{1-q} = \frac{\ln(e^{-q\Lambda_F})}{1-q} = \frac{q\ln(\tilde{P}/T) + \ln(\rho)}{1-q}. \tag{3}$$

Application of Renyi entropy for investigation of TM results is useful due to the following reasons. Firstly, Renyi entropy determines the degree to which the results of TM are non-equilibrium, so it accounts for the contribution of the initial distribution of the topic model. Secondly, topic models can be optimized based on finding the minimum of Renyi entropy. Thirdly, Renyi entropy, in contrast to Gibbs-Shannon entropy, allows to account for two different processes: decrease in Gibbs-Shannon entropy and increase in internal energy both of which occur with the growth of the number of topics. What follows from this is the existence of an area where these two processes counterbalance each other. In this area free energy and, correspondingly, Renyi entropy have the minimum values. Minimum of entropy corresponds to maximum of information of a topic model [Koltcov, 2018]. Hence, evaluation of the influence of hyper-parameters on the results of TM can be measured by means of Renyi entropy.

4 Description of data and computer experiments.

For our numerical experiments we used the following datasets:

- **Russian Dataset (from lenta.ru news agency)**: it consists of 8,630 news texts (containing 23,297 unique words) in Russian language, each of which is manually assigned with a class from a set of 10 topic classes. However, some of these topics are strongly correlated to each other. Therefore, the documents in this dataset can be represented by 7-10 topics.
- **English Dataset (the well-known '20 Newsgroups' dataset)**: 15,404 English news articles containing 50,948 unique words [1]. Each of the news items belongs to one or more of 20 topic groups. Since some of these topics can be unified, 14-20 topics can represent the documents of this dataset [Basu *et al.*, 2008].

In order to determine the influence of regularization on TM we investigated the following models: 1) pLSA model [Hofmann, 2001], a basic model with only one parameter - 'number of topics'; 2) LDA (GS) model [Griffiths and Steyvers, 2004], that can be considered a regularized extension of pLSA, where regularization is based on prior Dirichlet distributions with parameters α and β; 3) BigARTM model [Vorontsov and Potapenko, 2014] with smoothing/sparsing regularizers for matrix Φ (smooth/sparse phi) and matrix Θ (smooth/sparse theta), here termed sparse phi and sparse theta, respectively. These regularizers allow a user to obtain subsets of topics highly manifest in a small number of texts and/or words (sparsing effect), as well as subsets of topics relatively evenly distributed across all texts and words (smoothing effect). The parameter that controls the value of sparsing is a regularization coefficient termed τ. Additionally, we compared the results of the Renyi entropy approach for determining the 'optimal' number of topics with the results of HDP model. In our numerical experiments the number of topics T was varied in the range [2;50] in the increments of one topic. For LDA model hyper-parameters α and β were varied in the range [0.1;1] in the increments of 0.1. For BigARTM model we used the following values of τ: 0.01, 0.1, 1, and 10. For each topic model and for each dataset we calculated log-likelihood and Renyi entropy.

Let us note that computational efficiency of Renyi entropy approach turned out to be much higher than that of log-likelihood. For instance, calculation of Renyi entropy for the Russian dataset under variation of T in the range [2;50] in the increments of one took about 15 minutes, while calculation of log-likelihood for the same data took about nine hours. Such a great difference occurs because for Renyi entropy calculation it is enough to scan matrix Φ once, while for log-likelihood calculation one needs to multiply components of two large matrices (Φ and Θ). The purpose of our experiments was, firstly, to confirm that Renyi entropy allows us to determine the 'optimal' number of topics for the above datasets and to compare the results of this approach with results obtained by HDP model. Secondly, the purpose was to estimate the influence of hyper-parameters on results of TM and to specify which variant of regularization gives better results according to log-likelihood and Renyi entropy.

4.1 Optimal number of topics: HDP vs Renyi entropy in LDA (GS) and pLSA

To compare the results of HDP model, pLSA and LDA (GS), we calculate weights of topics for HDP model, and Renyi entropy for pLSA and LDA. HDP is a powerful model to cluster

[1] http://qwone.com/~jason/20Newsgroups/

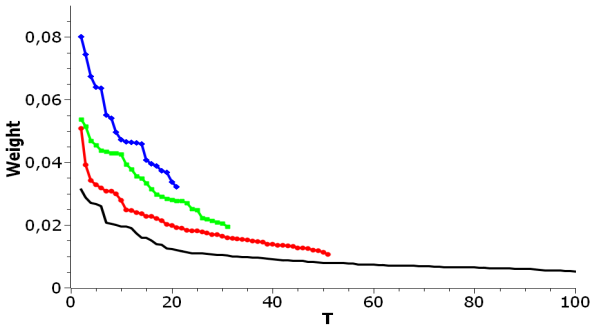

Figure 1: Distribution of weights over the number of topics T for HDP model (Russian dataset). TLT (100) – black, TLT (50) – red, TLT (30) – green, TLT (20) – blue.

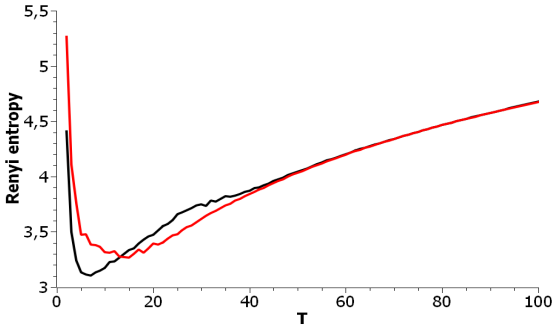

Figure 2: Renyi entropy distribution over the number of topics T (LDA). Russian dataset - black, English dataset - red.

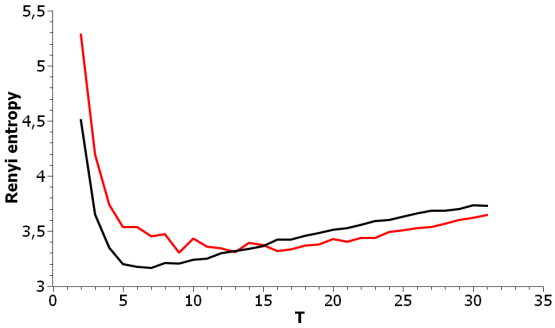

Figure 3: Renyi entropy distribution over the number of topics T (pLSA). Russian dataset - black, English dataset - red.

a collection of documents and inferring their topics without requiring the number of topics in advance [Teh *et al.*, 2006; Wang *et al.*, 2011]. Although this model is considered in the literature as non-parametric because it can model data with infinite number of topics, in real scenarios, users need to set a truncation on the allowed number of topics in the entire corpus. Since HDP returns the same number of topics as the top-level truncation that is set before, it is assumed that by discarding empty ones, the true number of topics can be obtained [Wang *et al.*, 2011]. In this experiment, we used the software adapted from [Yau *et al.*, 2014] to compute the weights of topics based on the obtained topic distribution. Fig. 1 plots together the outputs of four solutions of HDP model (Russian dataset) that differ by the values of top-level truncation parameter (TLT): 100, 50, 30, and 20. Each output is represented by a curve which sorts the weights of all inferred topics (whose number is always equal to TLT) in a descending order. The idea is to give the user an opportunity to cut off low-weight topics and to postulate that the "true" number of topics is equal to the number of high-weight topics. However, as can be seen, there is no clear threshold between high-weight and low-weight topics. The curves are monotone decreasing and do not allow to define the optimal number of topics. The same result was obtained for the English dataset. Moreover, we applied the method proposed by Wang and Blei [Wang and Blei, 2012] on both Russian and English corpora. This method proposes a truncation-free stochastic variational inference algorithm for HDP, which adapts the model complexity on the fly instead of requiring truncation values. For 100 runs, the method consistently inferred 28 topics on the Russian corpus and 24 topics for English corpus with default parameters.

Fig. 2 demonstrates Renyi entropy curves calculated according to equation (3) for LDA (GS) model. Here the number of topics was varied under fixed values of hyper-parameters: $\alpha = 0.5$, $\beta = 0.1$. Both curves have explicit minima of entropy, each of which is close to human mark-up. Fig. 3 shows Renyi entropy curves for pLSA model. As it can be seen, entropy curves for pLSA model and entropy curves for LDA (GS) model are very similar and the locations of minima are almost identical. However, Renyi entropy minimum for LDA model is more visible than for pLSA model.

4.2 Influence of hyper-parameters: pLSA vs LDA (GS) model

Let us discuss the influence of hyper-parameters α and β of LDA (GS) model on results of TM. Fig. 4 demonstrates dependence of log-likelihood on the number of topics for different values of α and β (Russian dataset). One can see that the increase in the values of hyper-parameters leads to the decrease in log-likelihood, which means that the model deteriorates as values of hyper-parameters increase. For $\alpha = \beta = 1$ we obtain the worst result for all numbers of topics. However, these curves do not allow us to determine simultaneously the optimal values of regularization parameters and the optimal number of topics. The behaviour of log-likelihood for English dataset is similar to that for Russian dataset and, therefore, we do not provide figure.

Fig. 5, 6 plot the curves of Renyi entropy for pLSA and LDA (GS) for different values of hyper-parameters. One can see that the increase in the values of hyper-parameters lifts the entire entropy curve, i.e., entropy increases on average. According to entropy approach the best model is the model with minimum entropy. It follows that the optimal models among the considered ones are pLSA and LDA (GS) with $\alpha = 0.1$, $\beta = 0.1$. Notice that minima of these optimal models coincide. Strong regularization ($\alpha = 1$, $\beta = 1$) leads not only to the growth of the entropy values on average but also to the horizontal shift of the minimum. One can conclude that the optimal values of hyper-parameters for LDA model with respect to Renyi entropy are $\alpha = 0.1$, $\beta = 0.1$. We can conclude that Renyi entropy approach allows us to determine both the optimal values of hyper-parameters and the optimal

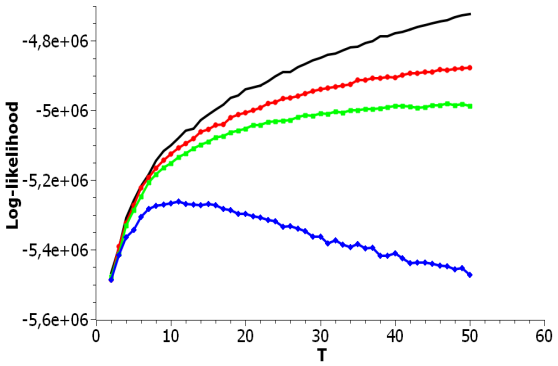

Figure 4: Log-likelihood distribution over T for different α and β (Russian dataset). pLSA – black, LDA (α=0.1, β=0.1) – red, LDA (α=0.5, β=0.1) – green, LDA (α=1, β=1) – blue.

number of topics, while log-likelihood metric allows us to determine the optimal values of hyper-parameters only.

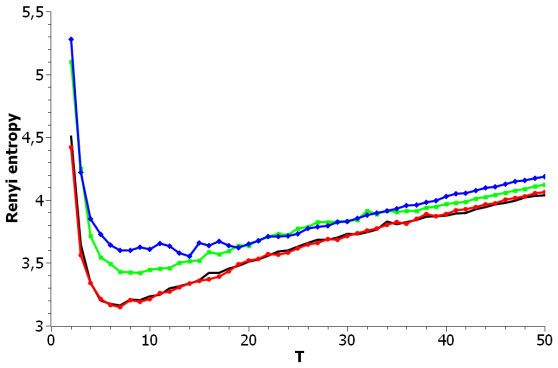

Figure 5: Renyi entropy distribution over T for different α and β (Russian dataset). pLSA – black, LDA (α=0.1, β=0.1) – red, LDA (α=0.5, β=0.1) – green, LDA (α=1, β=1) – blue.

4.3 Influence of regularization coefficients: BigARTM vs pLSA

We further discuss the influence of regularization parameters of BigARTM model on the results of TM. Here we consider sparsing regularizers for matrix Φ (sparse phi) and matrix Θ (sparse theta), where τ is regularization coefficient.

Fig. 7, 8 show the behavior of log-likelihood under variation of the number of topics for different values of regularization coefficients. Both figures show that the increase in regularization coefficient impairs the model, and the minimum value of regularization coefficient of BigARTM corresponds to pLSA. The same result is obtained for the English dataset. Let us note that the curve of log-likelihood does not allow us to understand what happens with TM if one changes regularization coefficient and the number of topics simultaneously.

Fig. 9, 10 plot Renyi entropy curves for BigARTM model run on the Russian dataset under variation of the number of topics for different values of regularization coefficient. One can see that the range of coefficients $[0.01; 1]$ gives small fluctuations in entropy minimum. In addition, these minima are located in the range [7;10] which corresponds to human mark-up for this dataset. However, values of regularization coefficient $\tau > 1$ lead to significant distortion of the Renyi

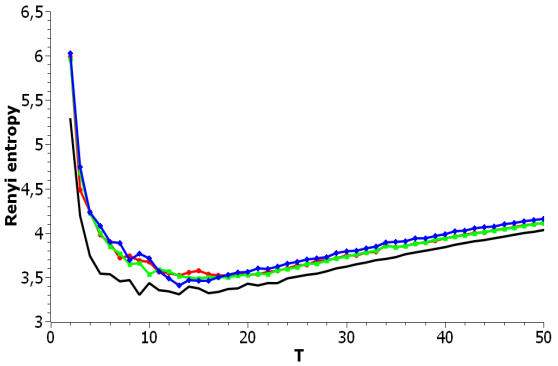

Figure 6: Renyi entropy distribution over T for different α and β (English dataset). pLSA – black, LDA (α=0.1, β=0.1) – red, LDA (α=0.5, β=0.1) – green, LDA (α=1, β=1) – blue.

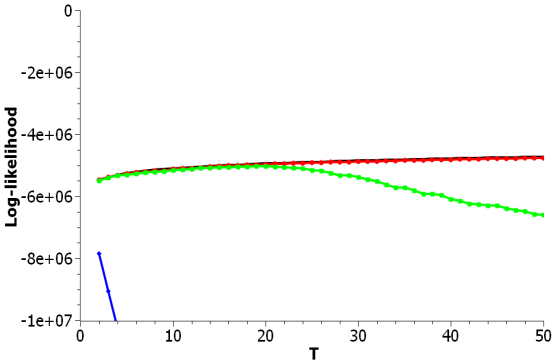

Figure 7: Log-likelihood distribution over T for different sparse phis (Russian dataset): 1. pLSA – black. 2. BigARTM sparse phi (τ=0.01) – red. 3. BigARTM sparse phi (τ=0.1) – green. 4. BigARTM sparse phi (τ=1) – blue.

entropy curve, i.e., to the lift of the entire curve and to the shift of the Renyi entropy minimum away from the known number of topics. This behavior is similar to that observed in fig. 5, 6 for hyper-parameters of classical LDA.

Likewise, the behavior of Renyi entropy for BigARTM on the English dataset (fig. 11, 12) is identical to that for the Russian dataset: the curve gets distorted when τ reaches the same value $\tau > 1$. Additionally, in both datasets the distortion introduced by regularizing phi is visibly larger than the effect of theta. Our experiments show the existence of a trade-off between model quality as determined by entropy, and regularization that allows to obtain e.g. sparse or smooth topics. In BigARTM, the smallest distortions are observed with the smallest τ which yields solutions close to the entirely unregularized model - pLSA. A similar result was obtained in [Apishev *et al.*, 2017], where pLSA was shown to perform better than any regularized BigARTM model, except the one with a dictionary-based regularizer. This was shown for the task of revealing ethnicity-related topics in social media texts by using coherence metric and human mark-up of topic interpretability.

5 Conclusion

We have proposed a method based on Renyi entropy for estimating the influence of model hyper-parameters and of regularization on the results of TM. This method was tested on

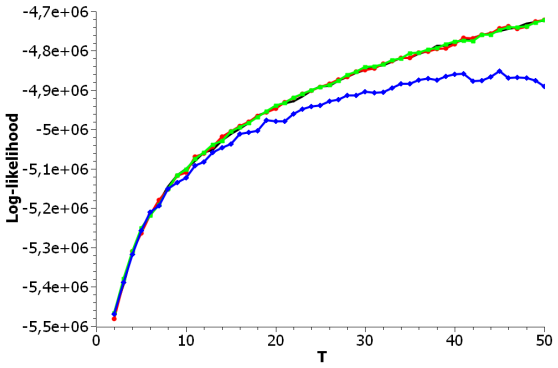

Figure 8: Log-likelihood distribution over T for different sparse thetas (Russian dataset): 1. pLSA – black. 2. BigARTM sparse theta (τ=0.01) – red. 3. BigARTM sparse theta (τ=0.1) – green. 4. BigARTM sparse theta (τ=1) – blue.

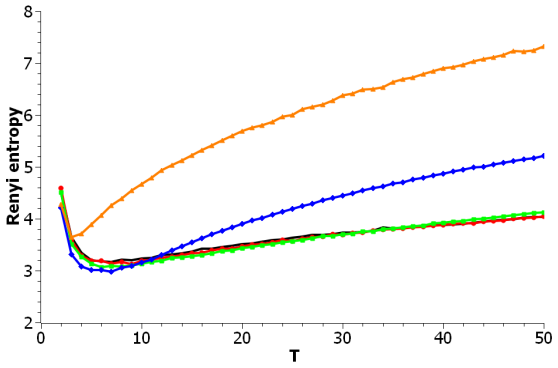

Figure 9: Renyi entropy distribution over T for different sparse phis (Russian dataset): 1. pLSA – black. 2. BigARTM sparse phi (τ=0.01) – red. 3. BigARTM sparse phi (τ=0.1) – green. 4. BigARTM sparse phi (τ=1) – blue. 5. BigARTM sparse phi (τ=10) – orange.

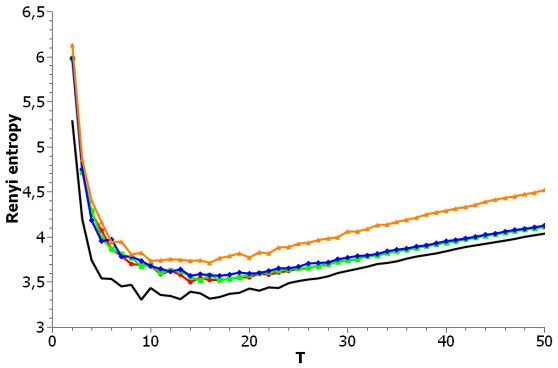

Figure 10: Renyi entropy distribution over T for different sparse thetas (Russian dataset): 1. pLSA – black. 2. BigARTM sparse theta (τ=0.01) – red. 3. BigARTM sparse theta (τ=0.1) – green. 4. BigARTM sparse theta (τ=1) – blue. 5. BigARTM spares theta (τ=10) – orange.

pLSA, LDA (Gibbs sampling) and BigARTM models. We demonstrated that higher levels of regularization and higher values of hyper-parameters lead to lower log-likelihood and higher entropy which is a clear sign of model deterioration. They also shift the minimum of Renyi entropy away from the optimal number of topics as determined by human-mark up, thus undermining the ability of this metric to indicate better

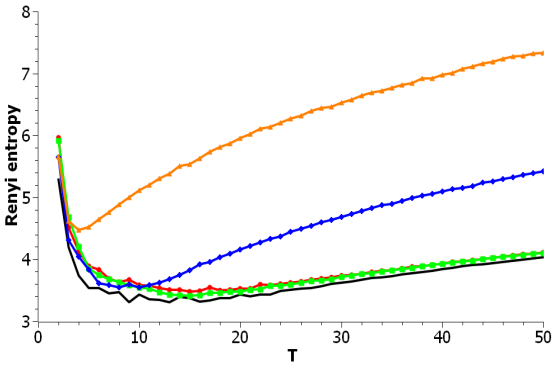

Figure 11: Renyi entropy distribution over T for different sparse phis (English dataset): 1. pLSA – black. 2. BigARTM sparse phi (τ=0.01) – red. 3. BigARTM sparse phi (τ=0.1) – green. 4. BigARTM sparse phi (τ=1) – blue. 5. BigARTM sparse phi (τ=10) – orange.

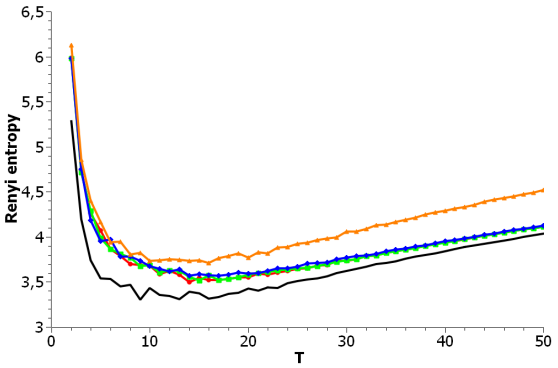

Figure 12: Renyi entropy distribution over T for different sparse thetas (English dataset): 1. pLSA – black. 2. BigARTM sparse theta (τ=0.01) – red. 3. BigARTM sparse theta (τ=0.1) – green. 4. BigARTM sparse theta (τ=1) – blue. 5. BigARTM sparse theta (τ=10) – orange.

solutions. However, since both metrics indicate the highest model quality there where the values of α, β and τ are low, Renyi entropy (unlike log-likelihood) may be used not only for finding the optima of those values, but also for finding an optimal number of topics, since it is in the range of low α, β and τ that Renyi entropy performs most accurately. In addition, calculation of Renyi entropy is simpler and faster than calculation of log-likelihood. Meanwhile, HDP does not provide clear thresholds to select the optimal number of topics. We conclude that Renyi entropy can be effectively used for estimating the influence of regularization coefficients and hyper-parameters on the results of TM, determining the optimal number of topics and estimating the effect of distortion under the condition of simultaneous change of multiple model parameters.

References

[Apishev *et al.*, 2017] Murat Apishev, Sergei Koltcov, Olessia Koltsova, Sergey Nikolenko, and Konstantin Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated texts. In *Advances in Computational Intelligence*, pages 169–184, 2017.

[Asuncion *et al.*, 2009] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 27–34, Arlington, Virginia, United States, 2009. AUAI Press.

[Basu *et al.*, 2008] S. Basu, I. Davidson, and K. Wagstaff, editors. *Constrained clustering : advances in algorithms, theory, and applications*. Chapman & Hall/CRC data mining and knowledge discovery series. Taylor & Francis Group Boca Raton, 1st. edition, 2008.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[De Waal and Barnard, 2008] A De Waal and E. Barnard. Evaluating topic models with stability. In *Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 79–84. PRASA, 2008.

[George and Doss, 2017] Clint P George and Hani Doss. Principled selection of hyperparameters in the latent dirichlet allocation model. *Journal of Machine Learning Research*, 18, 2017.

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

[Heinrich, 2004] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[Hofmann, 2001] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1/2):177–196, January 2001.

[Koltcov *et al.*, 2014] Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. Latent dirichlet allocation: Stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 161–165, New York, NY, USA, 2014. ACM.

[Koltcov, 2018] Sergei Koltcov. Application of rényi and tsallis entropies to topic modeling optimization. *Physica A: Statistical Mechanics and its Applications*, 512:1192 – 1204, 2018.

[Mora and Walczak, 2016] T. Mora and A. M. Walczak. Renyi entropy, abundance distribution and the equivalence of ensembles. 2016.

[Renyi, 1970] Alfred. Renyi. *Probability theory*. North-Holland Pub. Co Amsterdam, 1970.

[Teh *et al.*, 2006] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[Tkačik *et al.*, 2015] Gašper Tkačik, Thierry Mora, Olivier Marre, Dario Amodei, Stephanie E. Palmer, Michael J. Berry, and William Bialek. Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of the National Academy of Sciences*, 112(37):11508–11513, 2015.

[Tsallis, 2009] Constantino. Tsallis. *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World /*. Springer New York,, New York, NY :, 2009.

[Vorontsov and Potapenko, 2014] Konstantin Vorontsov and Anna Potapenko. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In *Analysis of Images, Social Networks and Texts*, Communications in Computer and Information Science. Springer International Publishing, 2014.

[Wallach *et al.*, 2009] Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 1973–1981, USA, 2009. Curran Associates Inc.

[Wang and Blei, 2012] Chong Wang and David M Blei. Truncation-free online variational inference for bayesian nonparametric models. In *Advances in neural information processing systems*, pages 413–421, 2012.

[Wang *et al.*, 2011] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 752–760. PMLR, 2011.

[Yau *et al.*, 2014] Chyi-Kwei Yau, Alan Porter, Nils Newman, and Arho Suominen. Clustering scientific documents with topic modeling. *Scientometrics*, 100(3):767–786, 2014.