TRUDY INSTITUTA MATEMATIKI I MEKHANIKI UrO RAN

Vol. 26 No. 3 **2020**

CONVERGENCE OF THE ALGORITHM OF ADDITIVE REGULARIZATION OF TOPIC MODELS

I. A. Irkhin, K. V. Vorontsov

The problem of probabilistic topic modeling is as follows. Given a collection of text documents, find the conditional distribution over topics for each document and the conditional distribution over words or terms for each topic. Log-likelihood maximization is used to solve this problem. The problem has generally an infinite set of solutions, being ill-posed according to Hadamard. In the framework of Additive Regularization of Topic Models (ARTM), a weighted sum of regularization criteria is added to the main log-likelihood criterion. The numerical method for solving this optimization problem is a kind of iterative EM-algorithm. In ARTM it is inferred in a quite general form for an arbitrary smooth regularizer, as well as for a linear combination of smooth regularizers. This paper studies the problem of convergence of the EM iterative process. Sufficient conditions are obtained for the convergence to a stationary point of the regularized log-likelihood. The constraints imposed on the regularizer are not too restrictive. We give their interpretations from the point of view of the practical implementation of the algorithm. A modification of the algorithm is proposed that improves the convergence without additional time and memory costs. Experiments on the news text collection have shown that our modification both accelerates the convergence and improves the value of the criterion to which it converges.

Keywords: natural language processing, probabilistic topic modeling, probabilistic latent semantic analysis (PLSA), latent Dirichlet allocation (LDA), additive regularization of topic models (ARTM), EM-algorithm, sufficient conditions for convergence.

MSC: 90C30, 68T50

DOI: 10.21538/0134-4889-2020-26-3-56-68

REFERENCES

1. Apishev M.A., Vorontsov K.V. Learning topic models with arbitrary loss. In: *Proceeding of the 26th Conf. of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association*, 2020, pp. 30–37. doi: 10.23919/FRUCT48808.2020.9087559 .
2. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 2003, vol. 3, pp. 993–1022.
3. Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society: Series B*, 1977, vol. 39, no. 1, pp. 1–38.
4. Frei O. I., Apishev M. A. Parallel Non-blocking Deterministic Algorithm for Online Topic Modeling. In: Ignatov D. et al. (eds), *Analysis of Images, Social Networks and Texts (AIST'2016)*, 2017, Communications in Computer and Information Science, vol. 661, Cham: Springer, pp. 132–144. doi: 10.1007/978-3-319-52920-2_13 .
5. Hofmann T. Probabilistic latent semantic indexing. In: *Proc. of the 22nd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval.* N Y: ACM, 1999, pp. 50–57. doi: 10.1145/312624.312649 .
6. Kochedykov D.A., Apishev M.A., Golitsyn L.V., Vorontsov K.V. Fast and modular regularized topic modelling. In: *Proc. of the 21st Conf. of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW), Helsinki, Finland, November 6–10, 2017*, 2017, pp. 182–193. doi: 10.23919/FRUCT.2017.8250181 .
7. Lang K. *20 newsgroups.* 2008. Data retrieved from the dataset's official website. Available at http://qwone.com/ jason/20Newsgroups/.
8. Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. In: *7th International Symposium Chinese Spoken Language Processing (ISCSLP).* 2010, pp. 224–228. doi: 10.1109/ISCSLP.2010.5684906 .
9. Tikhonov A.N., Arsenin V.Ya. *Solutions of ill-posed problems.* N Y etc.: John Wiley & Sons, 1977, 258 p. ISBN: 0-470-99124-0 .

10. Topsøe F. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 2000, vol. 46, no. 4, pp. 1602–1609. doi: 10.1109/18.850703 .
11. Vorontsov K.V. Additive regularization for topic models of text collections. *Dokl. Math.*, 2014, vol. 89, no. 3, pp. 301–304. doi: 10.1134/S1064562414020185 .
12. Vorontsov K.V., Frei O.I., Apishev M.A., Romov P.A., Suvorova M.A. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. In: Khachay M., Konstantinova N., Panchenko A., Ignatov D., Labunets V. (eds), *Analysis of Images, Social Networks and Texts*, 2015, Communications in Computer and Information Science, vol. 542, Cham: Springer, pp. 370–381. doi: 10.1007/978-3-319-26123-2_36 .
13. Vorontsov K.V., Potapenko A.A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. In: Ignatov D., Khachay M., Panchenko A., Konstantinova N., Yavorsky R. (eds), *Analysis of Images, Social Networks and Texts*, 2014, Communications in Computer and Information Science, vol. 436, Cham: Springer, pp. 29–46. doi: 10.1007/978-3-319-12580-0_3 .
14. Vorontsov K.V., Potapenko A.A. Additive regularization of topic models. *Machine Learning*, 2015, vol. 101, no. 1, pp. 303–323. doi: 10.1007/s10994-014-5476-6 .
15. Vorontsov K.V., Potapenko A.A., Plavin A.V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. In: Gammerman A., Vovk V., Papadopoulos H. (eds), *Statistical Learning and Data Sciences*, 2015, Lecture Notes in Computer Science, vol. 9047, Cham: Springer, pp. 193–202. doi: 10.1007/978-3-319-17091-6_14 .
16. Wallach H.M., Mimno D.M., McCallum A. Rethinking LDA: why priors matter. In: *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09).* Red Hook: Curran Associates, pp. 1973–1981.
17. Wu C.J. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 1983, vol. 11, no. 1, pp. 95–103. doi: 10.1214/aos/1176346060 .

Received July 20, 2020
Revised August 6, 2020
Accepted August 17, 2020

Funding Agency: The work was performed within the project “Text mining tools for big data” according to the program of the Competence Center of the National Technological Initiative “Center for Big Data Storage and Processing” supported by the Ministry of Science and Higher Education of the Russian Federation under the agreement between Moscow State University and the NTI Fund of August 15, 2019, no. 7/1251/2019. This work was also partially supported by the Russian Foundation for Basic Research (project no. 20-07-00936).

Konstantin Vyacheslavovich Vorontsov, Dr. Phys.-Math. Sci, Prof., Moscow Institute of Physics and Technology (State University), Dolgoprudny, 141701 Russia, e-mail: k.v.vorontsov@phystech.edu .

Il'ya Aleksandrovich Irkhin, doctoral student, Moscow Institute of Physics and Technology (State University), Dolgoprudny, 141701 Russia, e-mail: ilirhin@gmail.com .

I. A. Irkhin, K. V. Vorontsov. Convergence of the algorithm of additive regularization of topic models, *Trudy Instituta Matematiki i Mekhaniki URO RAN*, 2020, vol. 26, no. 3, pp. 56–68 .