MILES Machine Intelligence and Learning Systems

Université Paris-Dauphine / LAMSADE
https://www.lamsade.dauphine.fr/wp/miles/

Dynamic topic modeling: from text to physics

Duration: 4 to 6 months
Supervisor: Alexandre Allauzen (alexandre.allauzen@dauphine.psl.eu)
URL: https://allauzen.github.io/research/positions/

Context: Topic models are statistical tools for discovering the hidden structure in a collection of observations. The most famous model is known as LDA (for latent Dirichlet allocation) [1] and was investigated in the field of natural language processing (NLP) to analyse large datasets of texts. With the development of deep-learning in NLP (especially Variational Auto-encoders), new models have emerged to design new topic models like [5, 2], including dynamic models [3]. Topic models can also be applied to other kind of data. For instance in this recent paper [4], the authors proposed to explore observations of a turbulent channel flow to identify structures.

Outline: The goals of this internship is to explore new architectures of neural topic models for different kind of data and applications. A non-exhaustive list of possible topics is:

- Extending architectures (encoder/decoder) to cope with the peculiarities of new kind of data: modeling for instance turbulent flow fields requires to redefine how observations are "encoded" by the model.
- Introducing correlation between topics in the model can help to better explain complexe data. However it increases the complexity of the inference algorithm.
- How to define a topic models for dynamical generative process: for instance recent attempts proposed to use recurrent architectures to capture how patterns vary over time.

Organisation : Depending on your skills, the trade-off between the experimental and the theorical parts can be adapted. For the experiments we will rely on pytorch and existing data simulators. The internship can be extended with a funded PhD position.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[3] Adji B Dieng, Francisco JR Ruiz, and David M Blei. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*, 2019.

[4] M. Frihat, B. Podvin, L. Mathelin, Y. Fraigneau, and F. Yvon. Coherent structure identification in turbulent channel flow using latent dirichlet allocation, 2020.

[5] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *International Conference on Learning Representations (ICLR)*, 2017.