Text-as-data: mathematical challenges

During the last few years, much data has become available as digital textual data, e.g. bug reports, parliamentary proceedings, medical journals, court cases, log files, and newspapers. The need to analyse larger corpora increases the demand for methods to draw scientifically and statistically valid conclusions from huge corpora. In classical statistics, historically, not much focus has been put on text as data. The reasons are many. Textual data is often large and noisy and curating is often needed before analysis. In addition, many models and inference methods are computationally too costly for textual data, partly due to large corpora and partly due to model size.

The increased availability of large corpora has increased the need for scientific and statistical inference and predictions for textual data. This PhD project's goal is to advance inference for textual data, with a special focus on the mathematical and theoretical challenges of models for textual data.

Recently there have been several significant developments in building models within NLP and machine learning (ML). Neural networks with transformer architecture, such as BERT, GPT-3 and LaMBDA, have transformed the field of NLP. These breakthroughs are however highly focused on prediction performance. In many applications, such as in the social sciences and the humanities, the aim is different. Instead, the goal of the analysis is inference. For example, can we draw the conclusion that the discourse regarding immigration has changed before and after the refugee crisis?

This gives rise to the general motivating question of this project: How can we draw valid scientific conclusions from large textual corpora? This question in turn gives rise to several different challenges from both a practical perspective, e.g. a need for new models, and a mathematical/theoretical perspective as we obviously want rigorous results on new and existing models and inference from them. The particular challenges that will be the focus in this project are:

1. Can we make valid statistical tests from textual models? These models, such as e.g. topic models are often complex, so computations under the null hypothesis may be difficult. Can that be made? Alternatively, can new models be constructed such that they are mathematically analysable and still relevant for scientific inference?
2. How can we use deep neural networks (with transformer architectures such as BERT and GPT) in combination with other statistical methods to make inference for large textual corpora? Since transformer-based neural networks can estimate contextual word embeddings, this shows promise of being part of models that indeed are relevant and analysable. Hence, this class of models may provide partial solutions to the previous problem.
3. How well does inference for textual data, e.g. such as MCMC or Hamiltonian Monte, scale with the size of the corpus? If inference does not scale well, are there efficient ways to go around the problem without sacrificing too much statistical accuracy? Research questions like this fall partly within the field of mixing times of Markov chains, where many very elegant results are known, but very few that are applicable to MCMC on Bayesian models in practice. For statistical inference, it is of essential importance that the inference algorithm for the model as such converges, since otherwise apparent differences may be caused by non-convergence alone.
4. What measures should be used for evaluation of textual models and how can such evaluation be made efficiently for a huge corpus? An extra challenge here appears from the

fact that textual models are most likely very misspecified (i.e. they do not truly represent the data generating mechanism).