# Deep Discrete Latent Variable Models - Exercises

Wilker Aziz

September 16, 2020

## 1 Analysing Twitter Data

We have a stream of tweets spanning a certain period of time. The data are organised in daily batches of tweets, each batch containing exactly $N$ tweets, each tweet made of at most $C$ characters.

Quite a bit of pre-processing has already taken place:

- Images have been automatically replaced by a token `image:class` where *class* indicates one out of a finite number of categories meant to capture the most salient property of an image;
- The tweets in the collection are all predominantly expressed in English (as far as an automatic language detection software can tell us), but they do contain hashtags, the codes for images, short URLs, emojis, slang, and all sort of stuff you should expect from social media data, possibly including foreign words.
- We have one version of the data where we applied tokenization, stripped punctuation, and normalised URLs using some scheme that retains information about the nature of the content (e.g., *news agency*, *government agency*, *science*, etc.), we also applied a number of strategies to reduce vocabulary size down to some manageable constant $V_1$.
- A second version of the dataset has been segmented using an algorithm such as BPE encoding or sentencepiece, and no other form of text normalisation was employed. The vocabulary is some small enough constant $V_2$.

As there are too many days in the data set, if necessary, it's okay to imagine that days are grouped into overlapping sequences of size $T$. And, though these sequences are not i.i.d., do pretend they are.

### 1.1 Part I

**Context** We want to analyse *trends* in terms of topics that people tweet about. The focus of the analysis is to gain insight about the data we already have and for now we do not have any future predictive task in mind. Initially we don't

really know what people talk about, so we would be happy enough to annotate each daily batch with 1 of $K$ latent topics. To gain insight about the topics, we would be happy to see examples of tweets or some other means to 'label' the topics for their semantic content.

**Task** Design a *tractable temporal model* that can be used to study this dataset. Propose a factorisation of the model as well as its parameterisation, and do employ NN architectures. Be explicit about your parametric choices and mind the domain of their parameters. Present the objective for parameter estimation via gradient-based methods, and discuss how to obtain key quantities such as gradients or gradient estimates.

**Analysis** We are looking for insights into

- can we represent a topic by a list of keywords?
- how long do topics stay in the platform before going unmentioned?
- do topics re-emerge after going unmentioned?
- for topics that re-emerge, how long do they stay unmentioned?
- the second time a topic emerges, does it trend for a shorter period of time?

How can we answer these questions with a trained model?

## 1.2 Part II

**Context** We used the model of Part I to study the dataset. After that we selected a sample of daily batches for annotation by humans. Those days have now been annotated with a subset of $D$ possible topics which are known to be relevant throughout the period covered by the dataset. These topics are not mutually exclusive, and the annotation is not associated with any tweet in particular, but with a daily batch.

**Task** Adjust the temporal model to account for topics that are not mutually exclusive in a day and to learn from all of the data, that is, the labelled and the unlabelled part. Propose a factorisation of the model as well as its parameterisation, and do employ NN architectures. If you need approximate inference, first explain why that is the case, then go with variational inference and propose a factorisation of the inference model as well as its parameterisation. Be explicit abut your parametric choices and mind the domain of their parameters. Present the objective for parameter estimation via gradient-based methods, and discuss how to obtain key quantities such as gradients or gradient estimates. Make sure to account for all data points (observed and unobserved).

**Analysis** We are looking for insights into

- what is the distribution of number of topics per day?
- do certain topics co-occur above chance level?
- are there topics whose opposite trends correlate?

How can we answer these questions with a trained model?