Report: Experiments on α updates

Arijus Pleska

This report assesses some experiments performed on reproducing α parameters used in a generative data process. Note that the report is structures in the following sections: 1) defining the experiment settings; 2) assessing the experiment results; 3) settings some questions to be discussed during next meeting.

Experiment Settings

The intention of the carried experiments is to identify the optimal settings for the Metropolis–Hastings algorithm application. To start with, I have generated a synthetic corpus; the parameters used in the corpus generation will allow to assess the performance achieved in the experiments. The corpus generation parameters are set as follows:

- The number of topics: K = 2;
- The number of documents (time-slices): T = 20;
- The size of vocabulary: V = 10;
- The number of words per document t: $N_t \sim \text{Pois}(\lambda), \quad \lambda = 1000.$

Further, to consider the initial settings of α_k development over documents, α_0 is a sine curve and α_1 is a cosine curve; the corresponding *softmax* expressions of the curves are illustrated in Figure 1 below.

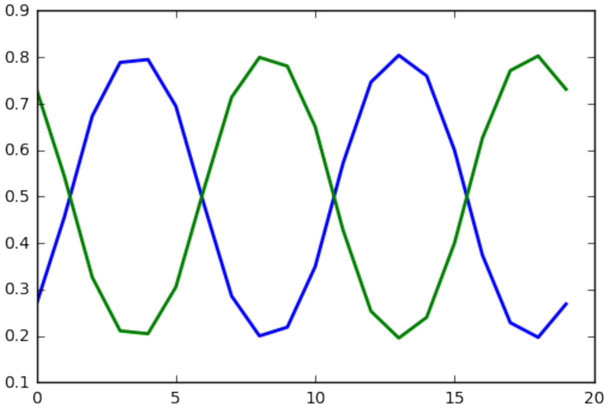

Figure 1: The values of μ used in the generative process.

Speaking of β, it was initially predefined and kept constant throughout the dynamic generative process; β is illustrated in Figure 2 below.

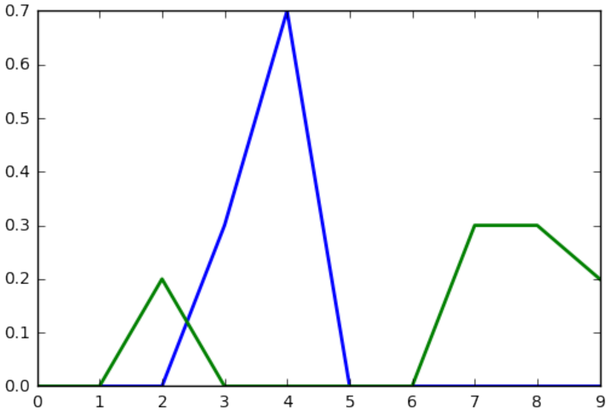

Figure 2: The values of β used in the generative process.

Note that the latter β values were applied to the autoregressive topic model for the α update experiments.

Current Stage

During the experiment, I have used the following settings:

- The synthetic data has been created by inducing the previously implemented dynamic topic modelling (DNT) generative process:
  - The number of documents: $|\mathrm{D}| \approx 6000$;
  - The size of the vocabulary: $|\mathrm{V}| \approx 2000$;
  - The number of words per document: $\mathrm{N}_d \approx 20, \quad \forall d \in \mathrm{D}$;
  - Instead of intensity values, it is assumed that the document dictionaries contain word counts. For example, $d_{111} = \{v_{20} : 15, v_{40} : 5\}$.
- The number of topics: $K = 10$;
- The number of time-slices: $T = 50$;
- The alpha at $t = 0$: $\alpha_0 \sim \mathcal{N}(\mu_0, \sigma_0^2 I), \quad \mu_0 = 0.1, \quad \sigma_0^2 = 0.2$;
- The alphas at $t > 0$: $\alpha_t \sim \mathcal{N}(\alpha_{t-1}, \sigma^2 I), \quad \sigma^2 = 0.1$;
- The candidate alphas: $\alpha_t' \sim \mathcal{N}(\alpha_t, \delta^2 I), \quad \delta^2 = 2$;

- The acceptance rate: $r_t = \min(1, p(\alpha'_t)/p(\alpha_t))$;
- The probability of the state: $p(\alpha_t) = p(\alpha_t|\alpha_{t-1}) \cdot p(\alpha_{t+1}|\alpha_t) \cdot \pi(\alpha_t)$, where π is a mapping to the mean parameterisation;

The rationale of the implementation follows the following principle: α_t is set to α'_t on the successful 'toss' based on r_t. Also, the variances are tuned to obtain $r_t \approx 30\%$.

Issues

My uncertainties with the proposed solution are the following:

- The estimation of $p(\alpha_t)$:
  - The third term of the expression, $\pi(\alpha_t)$, represents the topic distribution in documents in time-slice t;
  - The current model treats the vocabulary term distributions over the topics, β, to have same values; therefore, this term was omitted – it cancels out upon the estimation of r_t;
  - The first (and second) term $p(\alpha_t|\alpha_{t-1})$ is drawn from $\mathcal{N}(\alpha_{t-1}, \sigma^2 I)$.
- Since α_t is a vector, the initial r_t is a vector as well.