Report: The implementation of the Metropolis–Hastings (MH) algorithm to update the alphas

Arijus Pleska

This report is structured in two sections: an introduction to the current model's state; and a follow-up with the faced difficulties and knowledge gaps. Note that I have prioritised to implement the MH algorithm in order to establish a complete work-flow of the model. Even though I am uncertain about some concepts, hopefully, by having a report, it will be easier to discuss the issues during the project meetings.

Current Stage

During the experiment, I have used the following settings:

- The synthetic data has been created by 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.