Report: Experiments on α updates

Arijus Pleska

The purpose of this report is to assesses whether an auto-regressive topic model is capable to recover the α parameter used in the synthetic data generation process. This report is structured in the following order: at the start, the rationale of the used Bayesian methods are covered; then, the experiment settings are defined; in the next stage, we assess the experiment results; finally, the identified issues are outlined to be discussed during next meeting.

Preliminaries

In order to understand the experiment settings, it is necessary to be familiar with the auto-regressive and non auto-regressive α priors as well as the Metropolis–Hastings (M–H) algorithm; the expressions of the latter concepts are listed below:

1. The auto-regressive α prior:

$$p(\alpha_0, \ldots, \alpha_T) = p(\alpha_0) \prod_{t=1}^{T} p(\alpha_t|\alpha_{t-1}); \qquad p(\alpha_t) = f(\alpha_0; 0, \sigma_0^2 I), \quad t = 0;$$
$$p(\alpha_t) = f(\alpha_t; \alpha_{t-1}, \sigma^2 I), \quad t > 0.$$

2. The non auto-regressive α prior:

$$p(\alpha_0, \ldots, \alpha_T) = \prod_{t=0}^{T} p(\alpha_t); \qquad p(\alpha_t) = f(\alpha_0; 0, \sigma_0^2 I), \quad t \geq 0.$$

3. The rationale of the utilised M–H algorithm variation:

   For the start, familiarise with the following expressions:

$$\frac{A(x'|x)}{A(x|x')} = \frac{p(x'|x)}{p(x|x')} \cdot \frac{q(x'|x)}{q(x|x')} = \frac{p(x', x)}{p(x)} \cdot \frac{p(x')}{p(x, x')} = \frac{p(x')}{p(x)}; \qquad x' \sim q(x, \delta^2 I);$$
$$q = \mathcal{N} \Rightarrow q(x'|x) = q(x|x');$$

   where x is the current state, x' is the proposed state, A is the acceptance distribution, and q is the proposal distribution. Taking the previous results into account, the acceptance rate r is expressed as follows:

$$r = \min\left(1, \frac{p(x')}{p(x)}\right).$$

4. The application of the M–H algorithm to update α:

Note that the α values are updated independently; that is, the expression for the acceptance rate below is for a single α entry.

$$\frac{p(z,\alpha^{-tk},\alpha'_{t,k}|X)}{p(z,\alpha|X)} = \frac{p(X|z,\alpha^{-tk},\alpha'_{t,k})\cdot p(z|\alpha^{-tk},\alpha'_{t,k})\cdot p(\alpha^{-tk},\alpha'_{t,k})}{p(X)} \cdot \frac{p(X)}{p(X|z,\alpha)\cdot p(z|\alpha)\cdot p(\alpha)}$$
$$= \frac{\prod_{k=0}^{K}\left[\pi(\alpha'_t)_k^{z_{t,k}}\right]\cdot p(\alpha'_t|\alpha_{t-1})\cdot p(\alpha_{t+1}|\alpha'_t)}{\prod_{k=0}^{K}\left[\pi(\alpha_t)_k^{z_{t,k}}\right]\cdot p(\alpha_t|\alpha_{t-1})\cdot p(\alpha_{t+1}|\alpha_t)}; \qquad t>0, \quad t\neq T;$$

where π is the softmax function, $\alpha'_{t,k}\sim\mathcal{N}(\alpha_{t,k},\delta^2)$, and α^{-tk} denotes α without $\alpha'_{t,k}$; also, for the boundary cases $t=0$ and $t=T$, note that the $p(\alpha)$ term corresponds to $p(\alpha_0)\cdot p(\alpha_1|\alpha_0)$ and $p(\alpha_T|\alpha_{T-1})$ respectively. Further, for computational stability, the previous ratio would be computed in the log space as follows:

$$\log\left[\frac{p(z,\alpha^{-tk},\alpha'_{t,k}|X)}{p(z,\alpha|X)}\right] = \log\left[p(z,\alpha^{-tk},\alpha'_{t,k}|X)\right] - \log\left[p(z,\alpha|X)\right]; \quad \text{where}$$
$$\log\left[p(z,\alpha^{-tk},\alpha'_{t,k}|X)\right] = \sum_{k=0}^{K}\left[z_{t,k}\cdot\log\left[\pi(\alpha'_t)_k\right]\right] + \log\left[p(\alpha'_t|\alpha_{t-1})\right] + \log\left[p(\alpha_{t+1}|\alpha'_t)\right],$$
$$\log\left[p(z,\alpha|X)\right] = \sum_{k=0}^{K}\left[z_{t,k}\cdot\log\left[\pi(\alpha_t)_k\right]\right] + \log\left[p(\alpha_t|\alpha_{t-1})\right] + \log\left[p(\alpha_{t+1}|\alpha_t)\right].$$

Finally, the acceptance rate is calculated using the formula below.

$$r_{t,k} = \exp\left[\min\left(0, \log\left[\frac{p(z,\alpha^{-tk},\alpha'_{t,k}|X)}{p(z,\alpha|X)}\right]\right)\right].$$

The Experiment Settings

The intention of the carried experiments is to identify the optimal settings for the Metropolis–Hastings algorithm application. The rationale of the carried experiments is based on generating a corpus with pre-defined α changes. Based on the experiments, we will determine which techniques display higher performance in reproducing the pre-defined α fluctuations over time. To expand on the corpus generation settings, the parameters used are listed below:

- 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: $N_t \sim \text{Pois}(\lambda), \quad \lambda = 1000$.

Speaking of α_k development over time (documents), α_0 is a sine curve and α_1 is a cosine curve; the corresponding topic distributions over documents (i.e., $\theta = \text{softmax}(\alpha)$) are illustrated in Figure 1 below.

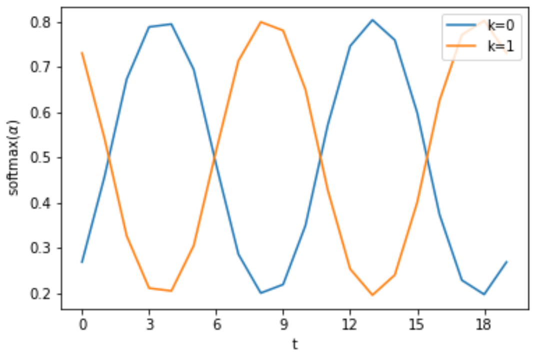

Figure 1: The values of softmax(α) used in the generative process.

Speaking of β, the parameter was initially pre-defined and kept constant throughout the dynamic generative process; β is illustrated in Figure 2 below.

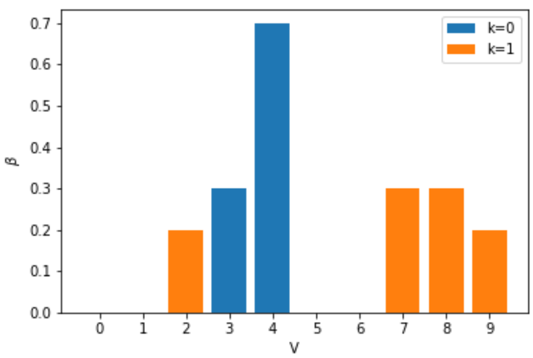

Figure 2: The values of β used in the generative process.

The experiment results

The first experiment is focused on discovering an optimal choice of the variances. Note that the first experiment is carried using the auto-regressive model; therefore, three different types of variances were considered: the 'basic' variance σ_0^2, the 'auto-regressive' variance σ^2, and the 'proposed' variance δ^2.

For the first experiment, δ^2 was kept constant and set to 1. Effectively, low δ^2 values suggest that the convergence of α is slow and stable, whereas for high values of δ^2 the convergence is faster and less stable. In both cases, with a high number of iterations, a low-error α value will be found. Therefore, we are focusing to tune only the σ_0^2 and σ^2 variances. Also, note that two visualisation of topic distribution over documents are provided: at 1000 iterations; and at 2000 iterations.

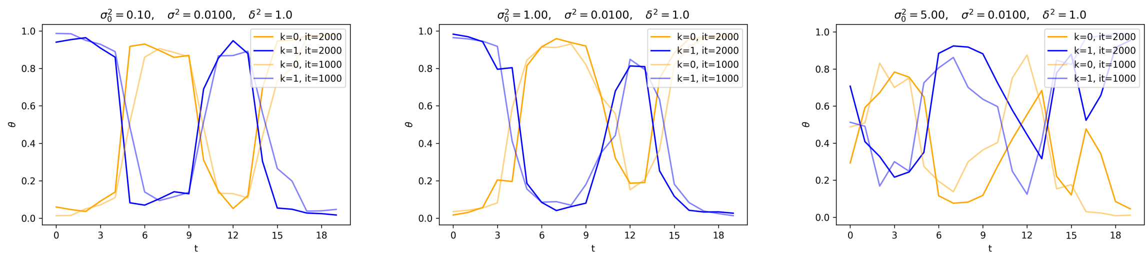

Figure 3: Fixed $\sigma^2 = 0.01$.

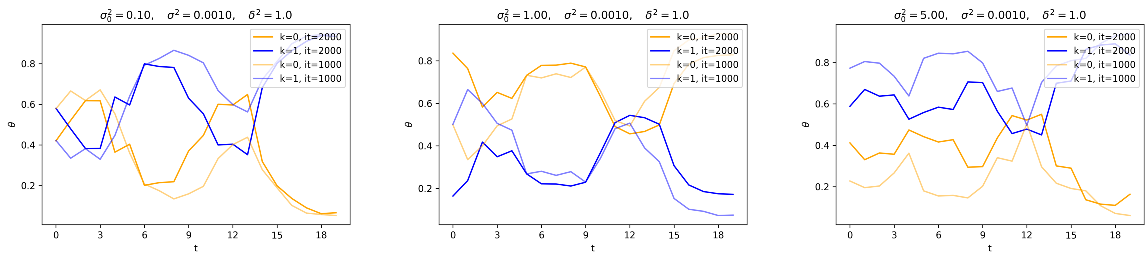

Figure 4: Fixed $\sigma^2 = 0.001$.

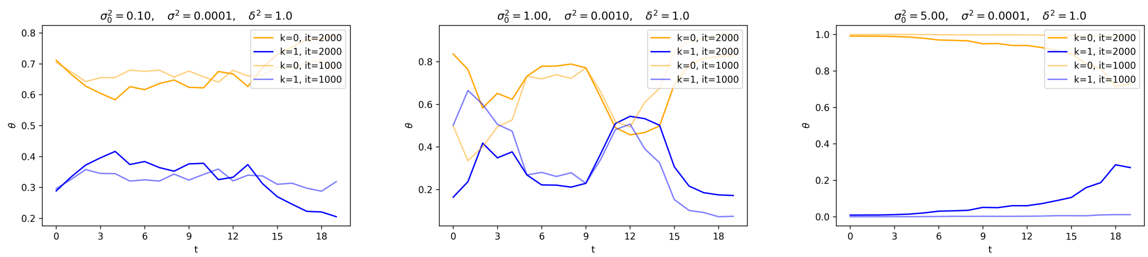

Figure 5: Fixed $\sigma^2 = 0.0001$.

The second experiment assesses the impact of the auto-regressive α update. For this reason, the α prior was switched to the non auto-regressive one. Note that in this case σ^2 has no impact; therefore, we provide illustrations by varying the σ_0^2 term. Again, $\delta^2 = 1$ was kept constant throughout the experiments. Also, note that we ran the model for 200 iterations. The resulting softmax(α) values are illustrated in Figure 6 below.

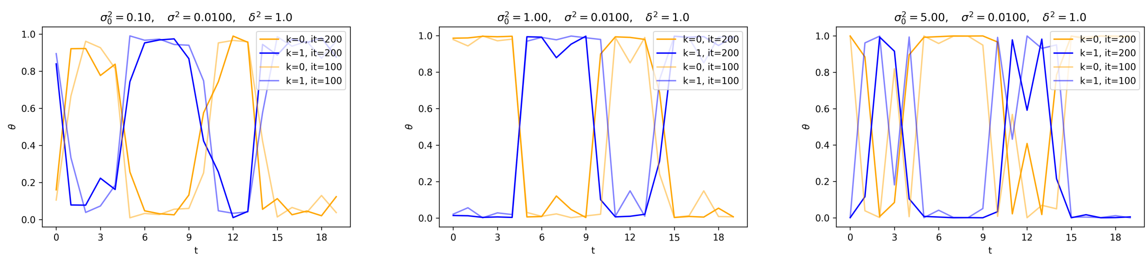

Figure 6: The non auto-regressive α update.

Questions

For the last section of this report, I have set some questions to be addressed during next meeting; these are listed below:

- During the derivation of the posterior given in Preliminaries Section, is the $\pi(\alpha'_t)_k$ term, i.e. $p(z|\alpha^{-tk}, \alpha'_{t,k})$, derived correctly? We did not have it explicitly expressed before.
- Should we consider the impact of tuning the model with the pre-defined β (the one used during the synthetic corpus generation)? The model is able to recover the topic assignments to documents even if β is initialised randomly. Also, do we have pre-defined β for the non synthetic data sets?
- What are the indications of the faster α converge displayed in the non auto-regressive model? Could it mean that the model implementation is faulty?