Scalable Generalized Dynamic Topic Models

Patrick Jähnichen[1] **Florian Wenzel**[1,2] **Marius Kloft**[2] **Stephan Mandt**[3]

[1]Humboldt-Universität zu Berlin, Germany [2]TU Kaiserslautern, Germany [3]Disney Research, Los Angeles, USA

Approximate marginalization

Following (Blei and Lafferty, 2006), we lower bound the intractable expectation in (8) by computing the first order Taylor approximation of the logarithm around an arbitrary location parameter $\zeta_{kt} > 0$,

$$\mathbb{E}_{p(\beta_t|u)}\left[\log \sum_v \exp(\beta_{kvt})\right] \leq \zeta_{kt}^{-1} \sum_v \exp\left(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} u_{kv.} + \frac{\tilde{K}_{tt}}{2}\right) + \log(\zeta_{kt}) - 1.$$

Updating the Taylor expansion location parameters.

In each iteration in our inference algorithm we optimize the location parameter of the Taylor expansion to achieve the tightest possible bound on true marginal likelihood (c.f. equation (9)). The derivative of the variational objective with respect to the location parameter of the Taylor expansion is

$$\frac{\partial \mathcal{L}}{\partial \zeta_{kt}} = \sum_n \phi_{tnk} \zeta_{kt}^{-1} \left(\zeta_{kt}^{-1} \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - 1\right).$$

Stetting the derivative to zero and solving for ζ_{kt} gives the update

$$\zeta_{kt} = \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right).$$

Derivation of the Variational Objective

Recall the variational objective

$$\mathcal{L}(\lambda, \phi, \mu, \Sigma) = \mathbb{E}_q[\log \tilde{p}(w|u, z) p(z|\theta) p(\theta) p(u)] - \mathbb{E}_q[\log q(\theta) q(z) q(u)].$$

The first term is

$$\begin{aligned}
&\mathbb{E}_q\left[\log \tilde{p}(w|z, u)\right] \\
&= \sum_{t,n,k} \mathbb{E}_q\left[z_{tnk} \log \tilde{p}(w_{tn}|z_{tn} = k, u)\right] \\
&= \sum_{t,n,k} \mathbb{E}_q[z_{tnk}] \left\{K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mathbb{E}_q[u_{k..}] w_{tn} - \zeta_{kt}^{-1} \sum_v \mathbb{E}_q\left[\exp\left(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} u_{kv.} + \frac{\tilde{K}_{tt}}{2}\right)\right] - \log(\zeta_{kt}) + 1\right\} \\
&= \sum_{t,n,k} \phi_{tnk} \Big\{K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mu_{k..} w_{tn} - \zeta_{kt}^{-1} \sum_v \exp\left(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \mu_{kv} + \frac{1}{2}(K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1} \Sigma_{kv} K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t} + \tilde{K}_{tt})\right) \\
&\quad - \log(\zeta_{kt}) + 1\Big\} \\
&= \sum_{t,n,k} \phi_{tnk} \left\{w_{tn.}^\top m_{k.t} - \zeta_{kt}^{-1} \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - \log(\zeta_{kt}) + 1\right\},
\end{aligned}$$

where $m_{kvt} = K_{t\hat{T}} K^{-1}_{\hat{T}\hat{T}} \mu_{kv}$ and $\Lambda_{kvt} = K_{t\hat{T}} K^{-1}_{\hat{T}\hat{T}} \Sigma_{kv} K^{-1}_{\hat{T}\hat{T}} K_{\hat{T}t}$.

The second term is

$$\begin{aligned}\mathbb{E}_q[\log p(z|\theta)] &= \sum_{t,n} \mathbb{E}_q[\log p(z_{tn}|\theta_t)] = \sum_{t,n,k} \phi_{tnk} \mathbb{E}_q[\log \theta_{tk}] \\ &= \sum_{t,n,k} \phi_{tnk} \left(\psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right)\right),\end{aligned}$$

where $\lambda_{t0} = \sum_k \lambda_{dk}$.

The negative KL terms are

$$\mathbb{E}_q[\log p(u) - \log q(u)] = -\sum_{k,v} \mathrm{KL}(q(u_{kv})||p(u_{kv})) \overset{c}{=} -\frac{1}{2} \sum_{k,v} \left(\mu_{kv} K^{-1}_{\hat{T}\hat{T}} \mu_{kv} + \mathrm{tr}(\Sigma_{kv} K^{-1}_{\hat{T}\hat{T}}) - \log|\Sigma_{kv}|\right).$$

and

$$\begin{aligned}\mathbb{E}_q[\log p(\theta) - \log q(\theta)] &= -\sum_t \mathrm{KL}(q(\theta_t)||p(\theta_t)) \\ &\overset{c}{=} \sum_{t,k} \left((\alpha_k - \lambda_{tk})(\psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right)) + \log\Gamma(\lambda_{tk})\right) - \Gamma(\lambda_{t0}).\end{aligned}$$

The entropy of $q(z)$ is

$$-\mathbb{E}_q[q(z)] = -\sum_{t,n,k} \phi_{tnk} \log \phi_{tnk}.$$

Finally, summing all terms gives the variational objective

$$\begin{aligned}\mathcal{L}(\lambda, \phi, \mu, \Sigma) = \sum_{t,n,k} \phi_{dnk} \Big\{ & w_{tn}^\top m_{k.t} - \zeta_{kt}^{-1} \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - \log(\zeta_{kt}) + 1 \\ & + \psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right) - \log\phi_{tnk} \Big\} - \frac{1}{2}\sum_{k,v}\left(\mu_{kv} K^{-1}_{\hat{T}\hat{T}} \mu_{kv} + \mathrm{tr}(\Sigma_{kv} K^{-1}_{\hat{T}\hat{T}}) - \log|\Sigma_{kv}|\right) \\ & + \sum_{t,k} \left((\alpha_k - \lambda_{tk})(\psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right)) + \log\Gamma(\lambda_{tk})\right) - \Gamma(\lambda_{t0}) + \mathrm{const}.\end{aligned}$$

SVI Updates

In the following we provide more details on how the the parameter updates are derived.

Updating the local variables of $q(z)$

The derivative of $\mathcal{L}$ w.r.t. ϕ_{tnk} is

$$\frac{\partial \mathcal{L}}{\partial \phi_{tnk}} = w_{tn}^\top m_{k.t} - \zeta_{kt}^{-1} \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt})\right) - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right) - \log\phi_{tnk}$$

Setting the derivative to zero leads to

$$\phi_{tnk} = \exp\left\{ w_{tn}^\top m_{k.t} - \zeta_{kt}^{-1} \sum_v \exp\left(m_{kvt} + \frac{1}{2}(\Lambda_{kvt} + \tilde{K}_{tt}) \right) - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right) \right\}.$$

Inserting the update of the previous update of ζ_{kt} this simplifies to

$$\begin{aligned} \phi_{tnk} &= \exp\left\{ w_{tn}^\top m_{k.t} - 1 - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right) \right\} \\ &\propto \exp\left\{ w_{tn}^\top m_{k.t} - \log(\zeta_{kt}) + \psi(\lambda_{tk}) - \psi\left(\lambda_{t0}\right) \right\}. \end{aligned}$$

The update for the parameter vector $\phi_{tn.}$ is obtained by renormalizing (such that $||\phi_{tn.}||_1 = 1$).

Updating the global variables

The standard Euclidean gradient of $\mathcal{L}$ with respect to the mean and covariance parameters of $q(u_{kv})$ is

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_{kv}} &= \Xi_{kv} - B_{kv} - K_{\hat{T}\hat{T}}^{-1}\mu_{kv}, \\ \frac{\partial \mathcal{L}}{\partial \Sigma_{kv}} &= -\frac{1}{2}C_{kv} + \frac{1}{2}\Sigma_{kv}^{-1} - \frac{1}{2}K_{\hat{T}\hat{T}}^{-1}, \end{aligned}$$

where $\Xi_{kv} = \sum_{t,n} \phi_{tnk} w_{tnv} K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t}$, $B_{kv} = \sum_{t,n} \zeta_{kt}^{-1}\phi_{tnk} \exp\left(m_{kvt} + \frac{\Lambda_{kvt} + \tilde{K}_{tt}}{2} \right) K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t}$ and $C_{kv} = \sum_{t,n} \zeta_{kt}^{-1}\phi_{tnk} \exp\left(m_{kvt} + \frac{\Lambda_{kvt} + \tilde{K}_{tt}}{2} \right) K_{\hat{T}\hat{T}}^{-1} K_{\hat{T}t} K_{t\hat{T}} K_{\hat{T}\hat{T}}^{-1}$.

We now consider the Gaussian distributions $q(u_{kv})$ in natural parametrization using $\eta_{kv}^{(1)} = S_{kv}^{-1}\mu_{kv}$ and $\eta_{kv}^{(2)} = -\frac{1}{2}S_{kv}^{-1}$. Applying formula (11) we obtain the natural gradient w.r.t. natural parameters,

$$\begin{aligned} \hat{\nabla}_{\eta_{kv}^{(1)}}\mathcal{L} &= \Xi_{kv} - B_{kv} - K_{\hat{T}\hat{T}}^{-1}\mu_{kv} - 2(-\frac{1}{2}C_{kv} + \frac{1}{2}\Sigma_{kv}^{-1} - \frac{1}{2}K_{\hat{T}\hat{T}}^{-1})\mu_{kv} \\ &= \Xi_{kv} + B_{kv} \circ (m_{kv} - 1) - \eta_{kv}^{(1)} \end{aligned}$$

and

$$\hat{\nabla}_{\eta_{kv}^{(2)}}\mathcal{L} = -\frac{1}{2}C_{kv} - \frac{1}{2}K_{\hat{T}\hat{T}}^{-1} - \eta_{kv}^{(2)}.$$

Note that m_{kv} as function of the natural parameters is

$$m_{kv} = K_{T\hat{T}} K_{\hat{T}\hat{T}}^{-1}\mu_{kv} = -\frac{1}{2}K_{T\hat{T}} K_{\hat{T}\hat{T}}^{-1}\left(\eta_{kv}^{(2)}\right)^{-1}\eta_{kv}^{(1)}.$$

Global td-idf score

To determine important words, we use an extension to the classic tf-idf scoring scheme. The score of a word

$$\text{score}(w) = \frac{n_w}{M}\ln\left(\frac{D}{n_{dw}}\right)$$

where M is the total amount of terms in the corpus, D is the number of documents, n_{dw} is the frequency of word w in document d and $n_w = \sum_d n_{dw}$.

References

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of ICML.*