Latent Dirichlet Allocation

Hamada Saleh

Contents

1 Optimization objective function: ELBO

The parameters of the variational distribution $q(z, \theta, \beta)$ (which approximates the true posterior) are optimized to maximize the Evidence Lower BOund (ELBO):

$$0 \geq \log p(w|\alpha, \eta) \geq \mathcal{L}(w, \phi, \gamma, \lambda)$$

Indeed, by maximizing the ELBO, we ensure that we are maximizing $\log p(w|\alpha, \eta)$, i.e. the logarithm of the true posterior, namely the log likelihood. In other words, we are performing **maximum likelihood estimation**. We want our probabilistic model to allocate high probability to documents that are actually seen in nature.

- K: number of topics
- W: vocabulary size

$$\begin{aligned}
\mathcal{L}(w, \Phi, \gamma, \lambda) &\triangleq \mathbb{E}_q[\log p(w, z, \theta, \beta|\alpha, \eta)] - \mathbb{E}_q[\log q(z, \theta, \beta)] \\
&= \mathbb{E}_q[\log p(\theta|\alpha, \eta)p(z|\alpha, \eta, \theta)p(\beta|\alpha, \eta, \theta, z)p(w|\alpha, \eta, \theta, z, \beta)] - \mathbb{E}_q[\log q(z)q(\theta)q(\beta)] \\
&= \mathbb{E}_q[\log p(\theta|\alpha)] + \mathbb{E}_q[\log p(z|\theta)] + \mathbb{E}_q[\log p(\beta|\eta)] + \mathbb{E}_q[\log p(w|\theta, z, \beta)] - \mathbb{E}_q[\log q(z)] \\
&- \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log q(\beta)] \\
&= \mathbb{E}_q[\log p(\theta_1, ..., \theta_D|\alpha)] + \mathbb{E}_q[\log p(z_1, ..., z_D|\theta)] + \mathbb{E}_q[\log p(\beta|\eta)] + \mathbb{E}_q[\log p(w_1, ..., w_D|\theta, z, \beta)] - \mathbb{E}_q[\log q(z_1, ..., z \\
&- \mathbb{E}_q[\log q(\theta_1, ..., \theta_D)] - \mathbb{E}_q[\log q(\beta)] \\
&= \mathbb{E}_q[\log \prod_d p(w_d|\theta, z, \beta)] + \mathbb{E}_q[\log \prod_d p(z_d|\theta)] - \mathbb{E}_q[\log \prod_d q(z_d)] + \mathbb{E}_q[\log \prod_d p(\theta_d|\alpha)] - \mathbb{E}_q[\log \prod_d q(\theta_d)] + \mathbb{E}_q[\log \\
\mathcal{L}(w, \Phi, \gamma, \lambda) &= \sum_d \{\mathbb{E}_q[\log p(w_d|\theta, z, \beta)] + \mathbb{E}_q[\log p(z_d|\theta)] - \mathbb{E}_q[\log q(z_d)] + \mathbb{E}_q[\log p(\theta_d|\alpha)] - \mathbb{E}_q[\log q(\theta_d)] + (\mathbb{E}_q[\log p(\beta|\eta)] - \mathbb{E}
\end{aligned}$$

1.1 Behind every expectation lies...

1.1.1 Topic

- $\beta \in \mathbb{R}^{K\text{x}W}$
- $\beta_k \sim Dir(\eta^W)$ where $\eta^W = \eta * \mathbf{1_W}$.
- $p(\beta|\eta) = p(\beta_1, ..., \beta_K|\eta) = \prod_k p(\beta_k|\eta)$
- $p(\beta_k|\eta) = \frac{1}{B(\boldsymbol{\eta})} \prod_{w=1}^W \beta_{kw}^{\eta-1}$

$$\begin{aligned}
\mathbb{E}_q[\log p(\beta|\eta)] &= \sum_{k=1}^K \mathbb{E}_q[\log p(\beta_k|\eta)] \\
&= \sum_{k=1}^K \mathbb{E}_q \left[\log \frac{1}{B(\boldsymbol{\eta})} \prod_{w=1}^W \beta_{kw}^{\eta-1} \right]
\end{aligned}$$

$$\boxed{\mathbb{E}_q[\log p(\beta|\eta)] = \sum_{k=1}^K \left\{ -W \log \Gamma(\eta) + \log \Gamma(W\eta) + \sum_{w=1}^W (\eta - 1)\mathbb{E}_q(\log \beta_{kw}) \right\}}$$

1.1.2 Topic proportion

- $\theta_d \in \mathbb{R}^K$
- $\theta_d \sim Dir(\alpha^K)$ where $\alpha^K = \alpha * \mathbf{1_K}$
- $p(\theta_d|\alpha) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \theta_{dk}^{\alpha-1}$

$$\begin{aligned}\mathbb{E}_q[\log p(\theta_d|\alpha)] &= \mathbb{E}_q\left[\log \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_{dk}^{\alpha-1}\right] \\ &= -\log B(\boldsymbol{\alpha}) + \sum_{k=1}^{K}(\alpha-1)\mathbb{E}_q(\log \theta_{dk})\end{aligned}$$

$$\boxed{\mathbb{E}_q[\log p(\theta_d|\alpha)] = \log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \sum_{k=1}^{K}(\alpha-1)\mathbb{E}_q(\log \theta_{dk})}$$

1.1.3 Topic index distribution

- $z_d \in \mathbb{R}^{N_d}$ (i.e. for a given document d with N_d words, each word is assigned a topic index)
- $z_{di} \sim \theta_d$
- $p(z_{di} = k|\theta_d) = \theta_{dk}$
- $p(z_{di}|\theta_d) = \theta_{dz_{di}}$

$$\begin{aligned}\mathbb{E}_q[\log p(z_d|\theta_d)] &= \mathbb{E}_q[\log \prod_{i=1}^{N_d} p(z_{di}|\theta_d)] \\ &= \sum_{i=1}^{N_d} \mathbb{E}_q[\log p(z_{di}|\theta_d)] \\ &= \sum_{i=1}^{N_d} \mathbb{E}_q[\log \theta_{dz_{di}}] \\ &= \sum_{i=1}^{N_d}\sum_{k=1}^{K} \log \theta_{dk} q(z_{di} = k) \\ &= \sum_{i=1}^{N_d}\sum_{k=1}^{K} \log \theta_{dk} \Phi_{dw_{di}k} \\ &= \sum_{i=1}^{N_d} \mathbb{E}_q[\mathbb{E}[\log \theta_{dz_{di}}|z_{di}]] \\ &= \sum_{i=1}^{N_d} \mathbb{E}_q\left[\sum_{k=1}^{K} q(z_{di} = k) \log \theta_{dk}\right] \\ &= \sum_{i=1}^{N_d}\sum_{k=1}^{K} \Phi_{dw_{di}k}\mathbb{E}_q[\log \theta_{dk}]\end{aligned}$$

$$\boxed{\mathbb{E}_q[\log p(z_d|\theta_d)] = \sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk}\mathbb{E}_q[\log \theta_{dk}]}$$

1.1.4 Word distribution

- $w_d \in \mathbb{R}^{N_d}$
- $w_{di} \sim \beta_{z_{di}}$

$$
\begin{aligned}
\mathbb{E}_q[\log p(w_d|\theta_d, z_d, \beta)] &= \sum_{i=1}^{N_d} \mathbb{E}_q[\log p(w_{di}|\theta_d, z_{di}, \beta)] \\
&= \sum_{i=1}^{N_d} \mathbb{E}_q[\log \beta_{z_{di} w_{di}}] \\
&= \sum_{i=1}^{N_d} \sum_{k=1}^{K} q(z_{di} = k) \log \beta_{k w_{di}} \\
&= \sum_{i=1}^{N_d} \sum_{k=1}^{K} \Phi_{d w_{di} k} \log \beta_{k w_{di}}
\end{aligned}
$$

$$
\boxed{\mathbb{E}_q[\log p(w_d|\theta_d, z_d, \beta)] = \sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk} \mathbb{E}_q[\log \beta_{kw}]}
$$

1.1.5 Variational topic

- $\beta_k, \lambda_k \in \mathbb{R}^W$
- $q(\beta_k) = \mathrm{Dir}(\beta_k; \lambda_k)$: cette distribution est une variable aléatoire
- $q\beta) = q(\beta_1, ..., \beta_K) = \prod_k q(\beta_k)$ par indépendance des $q(\beta_k)$ FLOU
- $q(\beta_k) = \frac{1}{B(\boldsymbol{\lambda_k})} \prod_{w=1}^{W} \beta_{kw}^{\lambda_{kw}-1}$

$$
\begin{aligned}
\mathbb{E}_q[\log q(\beta)] &= \sum_{k=1}^{K} \mathbb{E}_q[\log q(\beta_k)] \\
&= \sum_{k=1}^{K} \mathbb{E}_q \left[\log \frac{1}{B(\boldsymbol{\lambda_k})} \prod_{w=1}^{W} \beta_{kw}^{\lambda_{kw}-1} \right]
\end{aligned}
$$

$$
\boxed{\mathbb{E}_q[\log q(\beta)] = \sum_{k=1}^{K} \left\{ \log \Gamma(\sum_{w=1}^{W} \lambda_{kw}) - \sum_{w=1}^{W} \log \Gamma(\lambda_{kw}) + \sum_{w=1}^{W} (\lambda_{kw} - 1) \mathbb{E}_q(\log \beta_{kw}) \right\}}
$$

1.1.6 Variational topic distribution

- $\theta_d, \gamma_d \in \mathbb{R}^K$
- $q(\theta_d) = \mathrm{Dir}(\theta_d; \gamma_d)$: cette distribution est une variable aléatoire
- $q(\theta_d) = \frac{1}{B(\gamma_d)} \prod_{k=1}^{K} \theta_{dk}^{\gamma_{dk}-1}$

$$
\mathbb{E}_q[\log q(\theta_d)] = \mathbb{E}_q \left[\log \frac{1}{B(\gamma_d)} \prod_{k=1}^{K} \theta_{dk}^{\gamma_{dk}-1} \right]
$$

$$
\boxed{\mathbb{E}_q[\log q(\theta_d)] = \log \Gamma(\sum_{k=1}^{K} \gamma_{dk}) - \sum_{k=1}^{K} \log \Gamma(\gamma_{dk}) + \sum_{k=1}^{K} (\gamma_{dk} - 1) \mathbb{E}_q(\log \theta_{dk})}
$$

1.1.7 Variational topic index distribution

- $\Phi_d \in \mathbb{R}^{W\mathrm{x}K}$, $z_d \in \mathbb{R}^{N_d}$
- $q(z_{di} = k) = \Phi_{d w_{di} k}$ est un réel
- $q(z_{di}) = \Phi_{d w_{di} z_{di}}$ est une variable aléatoire

$$
\begin{aligned}
\mathbb{E}_q[\log q(z_d)] &= \sum_{i=1}^{N_d} \mathbb{E}_q[\log q(z_{di})] \\
&= \sum_{i=1}^{N_d} \mathbb{E}_q[\log \Phi_{dw_{di}z_{di}}] \\
&= \sum_{i=1}^{N_d} \sum_{k=1}^{K} \log \Phi_{dw_{di}z_{di}=k} q(z_{di}=k) \\
&= \sum_{i=1}^{N_d} \sum_{k=1}^{K} \log \Phi_{dw_{di}k} \Phi_{dw_{di}k}
\end{aligned}
$$

$$
\boxed{\mathbb{E}_q[\log q(z_d)] = \sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk} \log \Phi_{dwk}}
$$

1.2 Putting the pieces together

$$
\begin{aligned}
\mathcal{L}(w,\Phi,\gamma,\lambda) =& \sum_d \{\mathbb{E}_q[\log p(w_d|\theta_d, z_d, \beta)] + \mathbb{E}_q[\log p(z_d|\theta_d)] - \mathbb{E}_q[\log q(z_d)] \\
&+ \mathbb{E}_q[\log p(\theta_d|\alpha)] - \mathbb{E}_q[\log q(\theta_d)] + (\mathbb{E}_q[\log p(\beta|\eta)] - \mathbb{E}_q[\log q(\beta)])/D\} \\
=& \sum_d \{ \left(\sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk} \mathbb{E}_q[\log \beta_{kw}] \right) + \left(\sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk} \mathbb{E}_q[\log \theta_{dk}] \right) - \left(\sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk} \log \Phi_{dwk} \right) \\
&+ \left(\log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1) \mathbb{E}_q(\log \theta_{dk}) \right) - \left(\log \Gamma(\sum_{k=1}^{K} \gamma_{dk}) - \sum_{k=1}^{K} \log \Gamma(\gamma_{dk}) + \sum_{k=1}^{K} (\gamma_{dk} - 1) \mathbb{E}_q(\log \theta \right. \\
&+ \frac{1}{D} \sum_{k=1}^{K} \left(-W \log \Gamma(\eta) + \log \Gamma(W\eta) + \sum_{w=1}^{W} (\eta - 1) \mathbb{E}_q(\log \beta_{kw}) \right. \\
&\left. - \sum_{w=1}^{W} \log \Gamma(\lambda_{kw}) + \log \Gamma(\sum_{w=1}^{W} \lambda_{kw}) + \sum_{w=1}^{W} (\lambda_{kw} - 1) \mathbb{E}_q(\log \beta_{kw}) \right\} \\
=& \sum_d \left\{ \left(\sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk} (\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kw}] - \log \Phi_{dwk}) \right) \right. \\
&+ \left(- \log \Gamma(\sum_{k=1}^{K} \gamma_{dk}) + \sum_{k=1}^{K} \{(\alpha - \gamma_{dk}) \mathbb{E}_q(\log \theta_{dk}) + \log \Gamma(\gamma_{dk})\} - K \log \Gamma(\alpha) + \log \Gamma(K\alpha) \right) \\
&+ \frac{1}{D} \sum_{k=1}^{K} \left(- \log \Gamma(\sum_{w=1}^{W} \lambda_{kw}) + \sum_{w=1}^{W} \{(\eta - \lambda_{kw}) \mathbb{E}_q(\log \beta_{kw}) + \log \Gamma(\lambda_{kw})\} - W \log \Gamma(\eta) + \log \Gamma(W\eta) \right) \\
\mathcal{L}(w,\Phi,\gamma,\lambda) =& \sum_d \left\{ \left(\sum_{w=1}^{W} n_{dw} \sum_{k=1}^{K} \Phi_{dwk} (\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kw}] - \log \Phi_{dwk}) \right) \right. \\
&- \log \Gamma(\sum_{k=1}^{K} \gamma_{dk}) + \sum_{k=1}^{K} \{(\alpha - \gamma_{dk}) \mathbb{E}_q(\log \theta_{dk}) + \log \Gamma(\gamma_{dk})\} \\
&+ \frac{1}{D} \sum_{k=1}^{K} \left(- \log \Gamma(\sum_{w=1}^{W} \lambda_{kw}) + \sum_{w=1}^{W} \{(\eta - \lambda_{kw}) \mathbb{E}_q(\log \beta_{kw}) + \log \Gamma(\lambda_{kw})\} \right) \\
&\quad + \log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \frac{1}{D} (\log \Gamma(W\eta) - W \log \Gamma(\eta))
\end{aligned}
$$

□

On a donc:

$$
\mathcal{L}(w,\Phi,\gamma,\lambda) \triangleq \sum_d \ell(n_d, \Phi_d, \gamma_d, \lambda)
$$

Where $\ell(n_d, \Phi_d, \gamma_d, \lambda)$ is the contribution of document d to the ELBO.

2 Measure of fit: Perplexity

In order to assess whether a fitted topic model is relevant, the authors choose perplexity as a metric. Ideally, such a measure should be positively correlated with human judgment, such that when the topics produced by a model make sense to the human eye, high perplexity is output (and vice versa).

$$\text{perplexity}(n^{test}, \lambda, \alpha) \triangleq \exp\{-(\sum_i \log p(n_i^{test}|\alpha, \beta))/(\sum_{i,w} n_{iw}^{test})\}$$

Upper bound on perplexity

$$\text{perplexity} \leq \exp\{-(\sum_i \mathbb{E}_q[\log p(n_i^{test}, \theta_i, z_i|\alpha, \beta)] - \mathbb{E}_q[\log q(\theta_i, z_i)])/(\sum_{i,w} n_{i,w}^{test})\}$$

$$\begin{aligned}
\mathbb{E}_q[\log p(n_i^{test}, \theta_i, z_i|\alpha, \beta)] &= \mathbb{E}_q[\log p(\theta_i, z_i|\alpha, \beta)p(n_i^{test}|\alpha, \beta, \theta_i, z_i)] \\
&= \mathbb{E}_q[\log p(\theta_i, z_i|\alpha, \beta)\Pi_{w=1}^{W} p(w_{i,w}^{test}|\alpha, \beta, \theta_i, z_{i,w})^{n_{i,w}^{test}}] \\
&\boxed{= \mathbb{E}_q[\log p(\theta_i|\alpha)] + \mathbb{E}[\log p(z_i|\theta_i)] + \mathbb{E}[\log p(w_i^{test}|\theta_i, z_i, \beta)]} \\
&= \mathbb{E}_q[\log p(\theta_i, z_i|\alpha, \beta)] + \sum_{w=1}^{W} n_{i,w}^{test} \mathbb{E}_q[\log p(w_{i,w}^{test}|\alpha, \beta, \theta_i, z_i)] \\
&= \mathbb{E}_q[\log p(\theta_i|\alpha, \beta)p(z_i|\alpha, \beta, \theta_i)] + \sum_{w=1}^{W} n_{i,w}^{test} \mathbb{E}_q[\log p(w_{i,w}^{test}|\alpha, \beta, \theta_i, z_i)] \\
&= \mathbb{E}_q[\log p(\theta_i|\alpha)] + \mathbb{E}[\log p(z_i|\theta_i)] + \sum_{w=1}^{W} n_{i,w}^{test} \mathbb{E}_q[\log p(w_{i,w}^{test}|\theta_i, z_i, \beta)] \\
&= \log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1)\mathbb{E}_q(\log \theta_{ik}) \\
&+ \sum_{w=1}^{W} n_{iw} \sum_{k=1}^{K} \Phi_{iwk} \mathbb{E}_q[\log \theta_{ik}] + \sum_{w=1}^{W} n_{i,w}^{test} \sum_{k=1}^{K} \Phi_{iwk} \mathbb{E}_q[\log \beta_{kw}] \\
\mathbb{E}_q[\log p(n_i^{test}, \theta_i, z_i|\alpha, \beta)] &= \log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1)\{\Psi(\gamma_{ik}) - \Psi(\sum_{l=1}^{K} \gamma_{il})\} \\
&+ \sum_{w=1}^{W} n_{iw} \sum_{k=1}^{K} \Phi_{iwk} \{\Psi(\gamma_{ik}) - \Psi(\sum_{l=1}^{K} \gamma_{il})\}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_q[\log q(\theta_i, z_i)] &= \mathbb{E}_q[\log q(\theta_i)q(z_i)] \\
&= \mathbb{E}_q[\log q(\theta_i)] + \mathbb{E}_q[\log q(z_i)] \\
&= \log \Gamma(\sum_{k=1}^{K} \gamma_{ik}) - \sum_{k=1}^{K} \log \Gamma(\gamma_{ik}) + \sum_{k=1}^{K} (\gamma_{ik} - 1)\mathbb{E}_q(\log \theta_{ik}) + \sum_{w=1}^{W} n_{iw} \sum_{k=1}^{K} \Phi_{iwk} \log \Phi_{iwk}
\end{aligned}$$

$$\boxed{\mathbb{E}_q[\log q(\theta_i, z_i)] = \log \Gamma(\sum_{k=1}^{K} \gamma_{ik}) - \sum_{k=1}^{K} \log \Gamma(\gamma_{ik}) + \sum_{k=1}^{K} (\gamma_{ik} - 1)\{\Psi(\gamma_{ik}) - \Psi(\sum_{l=1}^{K} \gamma_{il})\} + \sum_{w=1}^{W} n_{iw} \sum_{k=1}^{K} \Phi_{iwk} \log \Phi_{iwk}}$$

3 The Newton-Raphson algorithm

Why do we need Newton-Raphson ? We want to find optimal values for α and η using Maximum Likelihood Estimation with respect to the ELBO. That is, we are looking for α^* and η^* that maximize the ELBO (because we want our model, which is encoding our assumptions about reality, to best fit said reality / the observed data, and one way of doing so is by choosing parameters that assign highest probability to what is actually observed; i.e. maximizing likelihood). How do we find such α^* and η^* ? Through the Newton-Raphson procedure !

How does Newton-Raphson work ? This algorithm is used for finding zeros of functions. We're interested in the ELBO's gradient's roots, because maximizers of that function can be found in that set of points.

For further exploring [3], [2].

3.1 Newton-Raphson algorithm in cubic time

Algorithm 1 Newton-Raphson algorithm

```
$\alpha \leftarrow \alpha_0$
while $|\Delta^{\text{Newton}}(\alpha)| > \epsilon$ do
    $\alpha^{(t+1)} \leftarrow \alpha^{(t)} + \Delta^{\text{Newton}}(\alpha^{(t)})$

end while
```

Δ^{Newton} is the *Newton step*. It is the **direction** that leads to a local minimum of the quadratic approximation of the gradient function.

$$\Delta^{\text{Newton}}(\alpha) \triangleq -H^{-1}(\alpha)g(\alpha)$$

Most of the time, the amplitude of the step taken in a given direction is adjusted with a learning rate / step size ρ. One often readjusts the step size at each iteration, in such a way that $\rho_t \xrightarrow{t\to\infty} 0$. While too small a learning rate slows down convergence, too high a step size often leads to divergence.

$$\Delta^{\text{Newton}}(\alpha, \rho_t) \triangleq -\rho_t H^{-1}(\alpha)g(\alpha)$$

Complexity $H(\alpha)$ is a square matrix of size K. Inverting it is $\mathcal{O}(K^3)$ complex, which is quite costly... But we can do better !

3.2 Newton-Raphson algorithm in linear time

Given a Hessian matrix H of a particular form, we can come up with a Newton-Raphson algorithm that scales linearly (p. 26 of [1]).

If the Hessian matrix is such that:

$$H = \text{diag}(h) + 1z1^T$$

Then:

$$H^{-1} = \text{diag}(h)^{-1} - \frac{\text{diag}(h)^{-1}11^T\text{diag}(h)^{-1}}{z^{-1} + \sum_j h_j^{-1}}$$

And:

$$\boxed{(H^{-1}g)_i = \frac{g_i - c}{h_i}}$$

Where

$$c = \frac{g^t h}{z^{-1} + \sum_{j=1}^{K} h_j^{-1}}$$

Complexity c only depends on the 2k values of g and h, which yields a Newton-Raphson algorithm in linear time.

3.3 Applying linear Newton-Raphson: How to update Dirichlet priors ?

When fitting a topic model, we must provide *a priori* values for α and η, the Dirichlet topic-document distribution and document-word distribution. The default values are respectively $\alpha = \frac{1}{K}$ and $\eta = \frac{1}{W}$. In other words, we start off with *symmetrical* priors, where each vector has identical values. Then we have two options. We can either leave them *as is*, or use the training data for updating their values. Doing so will result in *asymmetrical* Dirichlet (*posterior*) distributions. You can try out different updating settings and see what works best for you ! (e.g. don't update, update only α, update only η, update α and η) [4].

Given the particular ELBO at hand, we can use the above linear Newton-Raphson algorithm.

α update $\alpha = (\alpha_1, ..., \alpha_K)$.

Given ℓ the per-document contribution to the ELBO:

$$\begin{aligned}\ell(n_t, \gamma_t, \phi_t, \lambda) = {} & \mathbb{E}_q[\log p(w_d|\theta_d, z_d, \beta)] + \mathbb{E}_q[\log p(z_d|\theta_d)] \\ & - \mathbb{E}_q[\log q(z_d)] + \mathbb{E}_q[\log p(\theta_d|\alpha)] - \mathbb{E}_q[\log q(\theta_d)] \\ & + \frac{1}{D}\left(\mathbb{E}_q[\log p(\beta|\eta)] - \mathbb{E}_q[\log q(\beta)]\right)\end{aligned}$$

Let's compute its gradient relative to α

$$\nabla_\alpha \ell(n_t, \gamma_t, \phi_t, \lambda) = \nabla_\alpha \mathbb{E}_q[\log p(\theta_d|\alpha)]$$

$$\frac{\partial \ell}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j}\left(\log\Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i=1}^{K}\log\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i - 1)\mathbb{E}_q[\log\theta_{di}]\right)$$

$$\frac{\partial \ell}{\partial \alpha_j} = \Psi(\sum_{i=1}^{K}\alpha_i) - \Psi(\alpha_j) + (\alpha_j - 1)\mathbb{E}_q[\log\theta_{dj}]$$

Hence, the gradient:

$$\boxed{\nabla_\alpha \ell(\gamma_t) = \Psi(\sum_{i=1}^{K}\alpha_i) * 1_K - \Psi(\alpha) + (\alpha - 1_K) * \mathbb{E}_q[\log\theta_d]}$$

Furthermore:

$$\frac{\partial^2 \ell}{\partial \alpha_i \partial \alpha_j} = \Psi'(\sum_k \alpha_k) = \frac{\partial^2 \ell}{\partial \alpha_j \partial \alpha_i}$$

$$\frac{\partial^2 \ell}{\partial \alpha_j^2} = \Psi'(\sum_k \alpha_k) - \Psi'(\alpha_j) + \mathbb{E}_q[\log\theta_{dj}]$$

Hence, the Hessian matrix:

$$\boxed{H_\ell(\alpha) = \text{diag}(h_\alpha) + 1_K z_\alpha 1_K^T}$$

Where:

$$\boxed{\begin{aligned} h_\alpha &= \mathbb{E}_q[\log\theta_d] - \Psi'(\alpha) \\ z_\alpha &= \Psi'(\sum_k \alpha_k) \\ c &= \frac{\nabla_\alpha \ell(\gamma_t)^T h_\alpha}{z_\alpha^{-1} + h_\alpha^{-1} 1_K}\end{aligned}}$$

At the end of the day, **the update rule for** α is the following:

$$\boxed{\alpha \leftarrow \alpha - \rho_t \tilde{\alpha}(\gamma_t)}$$

Where:

$$\rho_t \triangleq (\tau_0 + t)^{-\kappa} \qquad \text{(step size)}$$
$$\tilde{\alpha}(\gamma_t) \triangleq H_\ell(\alpha)^{-1} \nabla_\alpha \ell(\gamma_t) \qquad \text{(step)}$$

η update $\eta = (\eta_1, ..., \eta_W)$

Given $\mathcal{L}$ the ELBO function:

$$\mathcal{L}(w, \Phi, \gamma, \lambda) = \sum_d \ell(n_d, \gamma_d, \phi_d, \lambda)$$

Let's compute its gradient relative to η:

$$\frac{\partial \mathcal{L}}{\partial \eta_j} = \frac{\partial}{\partial \eta_j} \mathbb{E}_q[\log p(\beta|\eta)]$$
$$\frac{\partial \mathcal{L}}{\partial \eta_j} = \sum_{k=1}^{K} \{-\Psi(\eta_j) + \Psi(\sum_{i=1}^{W} \eta_i) + (\eta_j - 1)\mathbb{E}_q[\log \beta_{kj}]\}$$

$$\boxed{\nabla_\eta \mathcal{L}(\lambda) = K * [\Psi(\sum_{j=1}^{W} \eta_j) * 1_W - \Psi(\eta)] + (\eta - 1_W) * 1_K^T \mathbb{E}_q[\log \beta]}$$

And its Hessian matrix:

$$\frac{\partial^2 \mathcal{L}}{\partial \eta_i \partial \eta_j} = K\Psi'(\sum_{l=1}^{W} \eta_l) = \frac{\partial^2 \mathcal{L}}{\partial \eta_j \partial \eta_i}$$

$$\frac{\partial^2 \mathcal{L}}{\partial^2 \eta_j} = -K\Psi'(\eta_j) + K\Psi'(\sum_{l=1}^{W} \eta_l) + \sum_{k=1}^{K} \mathbb{E}_q[\log \beta_{kj}]$$

$$\boxed{H_\mathcal{L}(\eta) = \text{diag}(h_\eta) + 1_W z_\eta 1_W^t}$$

Where

$$\boxed{\begin{aligned} h_\eta &= \sum_{k=1}^{K} \mathbb{E}_q[\log \beta_k] - K\Psi'(\sum_{l=1}^{W} \eta_l) \\ z_\eta &= K\Psi'(\sum_{l=1}^{W} \eta_l) \end{aligned}}$$

At the end of the day, **the update rule for** η is the following:

$$\boxed{\eta \leftarrow \eta - \rho_t \tilde{\eta}(\gamma_t)}$$

Where:

$$\rho_t \triangleq (\tau_0 + t)^{-\kappa} \qquad \text{(step size)}$$
$$\tilde{\eta}(\gamma_t) \triangleq H_\mathcal{L}(\eta)^{-1} \nabla_\eta \mathcal{L}(\lambda) \qquad \text{(step)}$$

Quid du déterminant
[1]

References

[1] David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 2003.

[2] Jonathan Huang. Maximum Likelihood Estimation of Dirichlet Distribution Parameters.

[3] Thomas P. Minka. Estimating a Dirichlet distribution. *Technical Report M.I.T.*, 2000.

[4] Shaheen Syed and Marco Spruit. Selecting Priors for Latent Dirichlet Allocation.