Variational Bayes

Elias Stengel-Eskin

August 8, 2017

1 The Problem

- Given our generative model (the adaptor grammar) and our data, we would like to find a posterior distribution:
- $P(Hypothesis \mid Evidence)$.
- $P(H \mid E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{\sum\limits_{\forall H} P(E|H)P(H)}$
- call the numerator of the fraction on the right the generative model.
  - composed of the product of the likelihood *of the hypothesis* ($P(E \mid H)$)
  - and the prior probability of the hypothesis ($P(H)$)
  - likelihood is not a probability but a measure of how well our hypothesis fits the data.
- denominator is the marginal likelihood of the data, $P(E)$.
- To find this, we need to marginalize out (sum over) all possible hypotheses.
- hypotheses are the range of values for all of the latent variables in our model, this summation is computationally intractable.
- In order to obtain a posterior, we are forced to use some approximate means of finding this denominator.
- Often, a sampling approach is used. However, sampling can be very slow to converge and is not easily parallelizable across multiple cores.
- The variational Bayesian approach, on the other hand, treats the problem of finding an appropriate marginal distribution as an optimization problem (this will be explained in more detail).

2 ELBO

2.1 Jensen's inequality

- Jensen's inequality states that for a convex function f and random variable X: $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

- We're using the logs of probabilities here, so our function is actually concave. As it turns out, Jensen's inequality works both ways, meaning we just switch the direction of the inequality: $\log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]$

2.2 Derivation

- looking to approximate is our denominator (marginal likelihood)

- One way we can do this is by using the Kullback Leibler (KL) divergence between this intractable integral and some variational distribution q.

  1. Let $q_\nu(Z)$ be a family of variational distributions with variational parameter ν.
  2. to get the marginal likelihood ($\log\ p(X \mid \Phi)$) we will take the KL divergence between $q_\nu(Z)$ and $p(Z \mid X, \Phi)$.
  3. KL divergence is given by Blei and Jordan (2006):

$$
\begin{aligned}
D_{KL}(q_\nu(Z) \mid\mid p(Z \mid X, \Phi)) &= \mathbb{E}_q[\log\ \frac{q_\nu(Z)}{p(Z \mid X, \Phi)}] \\
&= \mathbb{E}_q[\log\ q_\nu(Z) - \log\ p(Z \mid X, \Phi)] \\
&= \mathbb{E}_q[\log\ q_\nu(Z) - \log\ \frac{p(Z, X \mid \Phi)}{p(X \mid \Phi)}] \\
&= \mathbb{E}_q[\log\ q_\nu(Z) - (\log\ p(Z, X \mid \Phi) - \log\ p(X \mid \Phi))] \\
&= \mathbb{E}_q[\log\ q_\nu(Z)] - \mathbb{E}_q[\log\ p(Z, X \mid \Phi)] + \log\ p(X \mid \Phi)
\end{aligned}
$$

- If we think about what KL divergence represents, we can intuitively understand why it cannot be negative.

- From here, we can see how minimizing this equation is the same as maximizing the lower bound on $\log\ p(X \mid \Phi)$

$$
\begin{aligned}
0 &\leq \mathbb{E}_q[\log\ q_\nu(Z)] - \mathbb{E}_q[\log\ p(Z, X \mid \Phi)] + \log\ p(X \mid \Phi) \\
-\log\ p(X \mid \Phi) &\leq \mathbb{E}_q[\log\ q_\nu(Z)] - \mathbb{E}_q[\log\ p(Z, X \mid \Phi)] \\
\log\ p(X \mid \Phi) &\geq \mathbb{E}_q[\log\ p(Z, X \mid \Phi)] - \mathbb{E}_q[\log\ q_\nu(Z)]
\end{aligned}
$$

Another way to reach this same equation is by using Jensen's inequality.

- Consider the log marginal likelihood: $\log\ p(x \mid \Phi) = \log \sum_{\forall z \in \mathbf{Z}} p(x, z \mid \Phi)$

- The sum marginalizes out the hidden variables z in the joint probability distribution.

- Picking any variational distribution $q(z)$ we can multiply the by $\frac{q(z)}{q(z)}$
- $\log \sum_{\forall z \in \mathbf{Z}} (p(x, z \mid \Phi) * \frac{q(z)}{q(z)}) = \log \sum_{\forall z \in \mathbf{Z}} q(z) \frac{p(x,z|\Phi)}{q(z)}$
- Jensen's inequality implies $\log \sum_{\forall z \in \mathbf{Z}} q(z) \frac{p(x,z|\Phi)}{q(z)} \geq \sum_{\forall z \in \mathbf{Z}} q(z) \log \ \frac{p(x,z|\Phi)}{q(z)}$
- Recall $\log\ (\frac{x}{y}) = \log\ (x) - \log\ (y)$
- so this becomes:
$$\sum_{\forall z \in \mathbf{Z}} q(z) \log \ \frac{p(x, z \mid \Phi)}{q(z)} = \sum_{\forall z \in \mathbf{Z}} q(z)(\log p(x, z \mid \Phi) - \log q(z)) =$$
$$\sum_{\forall z \in \mathbf{Z}} q(z) \log p(x, z \mid \Phi) - \sum_{\forall z \in \mathbf{Z}} q(z) \log \ q(z) =$$
$$\sum_{\forall z \in \mathbf{Z}} q(z) \log p(x, z \mid \Phi) + \mathcal{H}(q)$$
- where $\mathcal{H}(q) = -\sum_{\forall z \in \mathbf{Z}} q(z) \log \ q(z)$ Blei et al. (2017)
- This first term is of the form of our expected value definition in 1.3, so our equation becomes: $\log\ p(X \mid \Phi) - KL(q(Z) \mid\mid p(Z \mid X, \Phi)) = \mathbb{E}_q[\log\ p(z, x \mid \Phi)] + H(q)$
- From this equation, we can see why minimizing KL divergence gives us the best possible value for our marginal likelihood.

3 Coordinate Ascent

The lower bound that we've derived above is a good start, but it means nothing if we cannot use it to converge on a solution. All it tells us right now is that given a certain variational distribution $q(Z)$ we can compute the lower bound on the log marginal likelihood. If we were incredibly lucky, we might guess this distribution on the first try and be done, but barring this astronomically unlikely event, we will need to update our variational distribution in some iterative manner.

3.1 Mean field approximation

The first problem with finding q is that it is a distribution over all latent variables Z. Since it is often the strong conditional relationships between latent variables that make inference intractable, we need to find a way to simplify the problem. Intuitively, q is an approximation (recall that L(q) is a lower bound, not an equality relation). Since we plan on improving q through updates, it turns out we can get away with breaking these interdependencies and proposing a variational distribution $q_i(z_i)$ for all $z_i \in Z$. Since we treat each variational distribution over a latent variable as independent, we get $q(Z) = \prod_i q_i(z_i)$

- very powerful assumption, because it allows us to optimize each variational distribution iteratively.

- That is to say, while holding all other variational distributions constant, we will find the variational parameters for $q_i(z_i)$ that maximize the marginal likelihood.
- recall $L(q) \geq \sum_{z_i \in Z} q(Z) \log\ p(X, Z|\Phi) + H(q)$

for concise notation:

- let $q_j = q_j(Z_j)$

rewrite:

$$L(q) \geq \sum_{z_i \in Z} \left(\prod_i q_i(z_i) \right) \log\ p(X, Z|\Phi) + H(q)$$

$$L(q) \geq \mathbb{E}_{\prod_i q_i(z_i)} \log\ p(X, Z|\Phi)[\log\ p(X, Z|\Phi)] + H(q)$$

Using the chain rule and expanding the entropy term, we can rewrite this expression as

$$\log\ p(X|\Phi) + \sum_{i=1}^{|Z|} \mathbb{E}_q[\log\ p(z_i|X, z_1, ..., z_{i-1}, \Phi)] - \sum_{i=1}^{|Z|} \mathbb{E}_q[\log\ q_{\nu_i}(z_i)]$$

- Since $p(X|\Phi)$ does not depend on the variational parameter ν_i we can ignore it
- (recall that this is a lower bound, not an exact equality, so generally $X \geq A + B \wedge A, B \geq 0 \Rightarrow X \geq B \wedge X \geq A$)
- Since Z is a set, we can reorder its elements any way we wish. If we reorder them each time so that z_i comes last, we can say:

$$\mathcal{L}_i = \mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q[\log\ q_{\nu_i}(z_i)]$$

3.2 Exponential family

We often use exponential family distributions
Note that for any exponential family distribution q_{ν_i},

$$q_{\nu_i}(z_i) = h(z_i) \exp \left\{ \nu_i^T z_i - a(\nu_i) \right\}$$

where $a(\nu_i)$ is the cumulant function, which for the first three derivatives is equivalent to the corresponding derivatives of the moment generating function. We can rewrite our equation using this form for $q_{\nu_i}(z_i)$:

$$\mathcal{L}_i = \mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q\left[\log\ \left(h(z_i)\exp\left\{\nu_i^T z_i - a(\nu_i)\right\}\right)\right]$$
$$= \mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q\left[\log\ (h(z_i))) + \nu_i^T z_i - a(\nu_i)\right]$$
$$= \mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q\big[\log\ (h(z_i)))\big] - \mathbb{E}_q[\nu_i^T z_i] + \mathbb{E}_q[a(\nu_i)]$$
$$= \mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q\big[\log\ (h(z_i)))\big] - \nu_i^T a'(\nu_i) + a(\nu_i)$$

Note that $\mathbb{E}_q[\nu_i^T z_i] = \nu_i^T a'(\nu_i)$ becomes since $\mathbb{E}_q(z_i) = a'(\nu_i)$ and ν_i^T factors out as a constant when taking the expectation with respect to q.

- main premise of variational inference is to cast the intractable calculation of the posterior as an optimization problem
- In most optimization problems, there are two general steps
  1. computing an objective function which will allow us to
  2. optimize the function by adjusting the parameters
- So far, we have an objective function (our lower bound $\mathcal{L}(q)$ on the marginal likelihood
- Setting the first derivative to 0 and solving allows us to obtain a point where the slope of the function has leveled out
- With a strictly convex function, this will be the global maximum. However, in most real-life scenarios, the objective function is not strictly convex
- This means that the result could very well be a **local** maximum
- Converging on local maxima is one of the risks run when using variational inference
- In the general case, taking the derivative of the objective function can be costly
  - especially since it must be done every time we want to update our parameters and for every parameter in the factorized representation of the variational distribution.
- However, using exponential family random variables will allow us to leverage some convenient mathematical facts and avoid this computation.

We are trying to optimize the function by adjusting the variational parameters, so we take the partial derivative of our function with respect to ν_i:

$$\frac{\delta}{\delta\nu_i}\mathcal{L}_i = \frac{\delta}{\delta\nu_i}\left(\mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q\big[\log\ (h(z_i)))\big] - \nu_i^T a'(\nu_i) + a(\nu_i)\right)$$
$$= \frac{\delta}{\delta\nu_i}\left(\mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q[\log\ h(z_i))]\right) - \left(\nu_i^T a''(\nu_i) + a''(\nu_i)\right) + a''(\nu_i)$$
$$= \frac{\delta}{\delta\nu_i}\left(\mathbb{E}_q[\log\ p(z_i|Z_{-i}, X, \Phi)] - \mathbb{E}_q[\log\ h(z_i))]\right) - \nu_i^T a''(\nu_i)$$

Setting this to 0 we get:

$$\begin{aligned} 0 &= \frac{\delta}{\delta\nu_i}\left(\mathbb{E}_q[\log\ p(z_i|Z_{-i},X,\Phi)] - \mathbb{E}_q[\log\ h(z_i))]\right) - \nu_i^T a''(\nu_i) \\ \nu_i^T a''(\nu_i) &= \frac{\delta}{\delta\nu_i}\left(\mathbb{E}_q[\log\ p(z_i|Z_{-i},X,\Phi)] - \mathbb{E}_q[\log\ h(z_i))]\right) \\ \nu_i &= \left(\frac{\delta}{\delta\nu_i}\mathbb{E}_q[\log\ p(z_i|Z_{-i},X,\Phi)] - \frac{\delta}{\delta\nu_i}\mathbb{E}_q[\log\ h(z_i))]\right)\left(a''(\nu_i)\right)^{-1} \end{aligned}$$

Recall that for this to work, q must be in the exponential family. Similarly, if $p(z_i|Z_{-i},X,\Phi)$ is a member of the exponential family, it can be rewritten:

$$p(z_i|Z_{-i},X,\Phi) = h(z_i)\exp\left\{g_i(Z_{-i},X,\Phi)^T z_i - a\left(g_i(Z_{-i},X,\Phi)\right)\right\}$$

where $g_i(Z_{-i},X,\Phi)$ is the natural parameter of distribution p (for exponential family distributions, this natural parameter has already been defined). Plugging this in for $p(z_i|Z_{-i},X,\Phi)$ (first in the expected value alone, for the sake of readability) and taking the derivative we get:

$$\begin{aligned} \mathbb{E}_q[\log\ p(z_i|Z_{-i},X,\Phi)] &= \mathbb{E}_q[\log\ h(z_i)] + \mathbb{E}_q[g_i(Z_{-i},X,\Phi)]^T a'(\nu_i) - \mathbb{E}_q\big[a\left(g_i(Z_{-i},X,\Phi)\right)\big] \\ \frac{\delta}{\delta\nu_i}\mathbb{E}_q[p(z_i|Z_{-i},X,\Phi)] &= \frac{\delta}{\delta\nu_i}\mathbb{E}_q[\log\ h(z_i)] + \left(\frac{\delta}{\delta\nu_i}\left(\mathbb{E}_q[g_i(Z_{-i},X,\Phi)]^T\right)a'(\nu_i) + \mathbb{E}_q[g_i(Z_{-i},X,\Phi)]^T a''(\nu_i)\right) \\ &\quad - \frac{\delta}{\delta\nu_i}\mathbb{E}_q\big[a\left(g_i(Z_{-i},X,\Phi)\right)\big] \\ &= \frac{\delta}{\delta\nu_i}\mathbb{E}_q[\log\ h(z_i)] + \mathbb{E}_q[g_i(Z_{-i},X,\Phi)]^T a''(\nu_i) \end{aligned}$$

Note that many of the expectations drop out; when differentiating with respect to one variable, if a function or value is not written in terms of that variable, then it is treated as a constant and drops out to 0. Substituting $\frac{\delta}{\delta\nu_i}\left(\mathbb{E}_q[\log\ p(z_i|Z_{-i},X,\Phi)]\right)$ for this in our first differentiation, we get:

$$\begin{aligned} \nu_i &= \left(\frac{\delta}{\delta\nu_i}\mathbb{E}_q[\log\ h(z_i)] + \mathbb{E}_q[g_i(Z_{-i},X,\Phi)]^T a''(\nu_i) - \frac{\delta}{\delta\nu_i}\mathbb{E}_q[\log\ h(z_i))]\right)\left(a''(\nu_i)\right)^{-1} \\ &= \left(\mathbb{E}_q[g_i(Z_{-i},X,\Phi)]^T a''(\nu_i)\right)\left(a''(\nu_i)\right)^{-1} \\ &= \mathbb{E}_q[g_i(Z_{-i},X,\Phi)] \end{aligned}$$

So the optimal value (when the derivative is 0) of $\nu_i = \mathbb{E}_q[g_i(Z_{-i},X,\Phi)]$. Since we have a closed form for the natural parameters of exponential family equations, we can compute this expectation without having to do the costly differentiation normally required to obtain the gradient of a function.

Now we need to obtain a closed-form expression for $\mathbb{E}_q[g_i(Z_{-i},X,\Phi)]$. Firstly, let Φ be composed of 2 parts ϕ_1 and ϕ_2, where ϕ_1 is the number of observations contributed by the prior, and ϕ_2 corresponds to the total effect of the observations on the sufficient statistic.

Because of the factorization and exponential family assumptions we made earlier, we can say that for z_i we have a conjugate prior distribution:

$$\begin{aligned} P_\pi(z_i|\phi_1,\phi_2) &= f(\phi_1,\phi_2)\exp\left\{z_i^T\phi_1 - \phi_2 A(z_i)\right\} \\ &= f(\phi_1,\phi_2)g(z_i)^{\phi_2}\exp\left\{z_i^T\phi_1\right\} \\ &\propto g(z_i)^{\phi_2}\exp\left\{z_i^T\phi_1\right\} \end{aligned}$$

where $A(z_i)$ is the cumulant function, and $f(\phi_1,\phi_2)$ is a normalizing constant.
Assuming the posterior over data and local hidden variables $P(X, Z_{-i} \mid z_i)$ is also in the exponential family and factorizes, we can say that for one data point x_n

$$\begin{aligned} P(x_n, z_n \mid z_i) &= h(x_n,z_n)g(z_i)\exp\left\{z_i^T t(x_n,z_n)\right\} \\ \Rightarrow P(X, Z_{-i} \mid z_i) &= \prod_{z_n\in Z_{-i}} h(x_n,z_n)g(z_i)^N\exp\left\{z_i^T t(x_n,z_n)\right\} \end{aligned}$$

where $t(x_n, z_n)$ is the sufficient statistic, which in most cases is simply the count of occurrences of the parameters. By Bayes' rule, the distribution over the selected hidden variable rewrites as

$$\begin{aligned} P(z_i|X,Z_{-i},\Phi) &\propto P(X,Z_{-i}|z_i)P(z_i|\Phi) \\ = \prod_{z_n\in Z_{-i}} h(x_n,z_n)g(z_i)\exp\left\{z_i^T t(X,Z_{-i})\right\}&g(z_i)^{\phi_2}\exp\left\{z_i^T\phi_1\right\} \\ \propto g(z_i)^N\exp\left\{z_i^T\sum_{z_n\in Z_{-i}} t(x_n,z_n)\right\}&g(z_i)^{\phi_2}\exp\left\{z_i^T\phi_1\right\} \\ \propto g(z_i)^{N+\phi_2}\exp&\left\{z_i^T\left(\phi_1+\sum_{z_n\in Z_{-i}} t(x_n,z_n)\right)\right\} \end{aligned}$$

This shows that the posterior $P(z_i \mid X, Z_{-i}, \Phi)$ has the same form as the prior (which is the definition of conjugacy) and has parameters:

$$P(z_i|X,Z_{-i},\Phi) = P_\pi\big(z_i|\phi_1+\sum_{z_n\in Z_{-i}} t(x_n,z_n), \phi_2+N\big)$$

This means that the natural parameters of the posterior distribution on global hidden variables has the natural parameters $\phi_1 + \sum_{z_n\in Z_{-i}} t(x_n, z_n)$ and $\phi_2 + N$, giving us a closed form for our expectation in equation 9:

$$\mathbb{E}_q[g_i(Z_{-i},X,\Phi)] = \mathbb{E}_q\begin{bmatrix}\phi_1+\sum_{z_n\in Z_{-i}} t(x_n,z_n) \\ \phi_2+N\end{bmatrix}$$

Hoffman et al. (2013)

4 Example: PCFGs

Note: much of this section uses Kurihara and Sato (2004) as a reference ~~Kurihara and Sato (2004).~~

Time to apply this to real model. Let's take something we're all familiar with: PCFGs
Recall a PCFG is a 5-tuple:

$$\{N, T, R, S, \Theta\}$$

where N are nonterminals, T are terminals, R are rules, S is start, and Θ are rule probabilities.
Additionally, let:

- $d = \{d_1, d_2, ..., d_n\}$ be the derivations of
- $x = \{x_1, x_2, ..., x_n\}$ the sentences in the corpus.
- Φ be our model parameters

Assuming that for each nonterminal, the rule weights are drawn from a Dirichlet distribution parametrized by α_A.
Writing this down as we did with our generic model before, we get

$$p(d, \Theta|x, \Phi) = \frac{p(x|d, \Theta, \Phi)p(\Theta|\Phi)}{\sum\limits_{\forall \Theta, d} p(x|d, \Theta, \Phi)p(\Theta|\Phi)}$$

Define variational distribution

$$q(d, \Theta|x) = q(d|x)q(\Theta|x) = \Big(\prod_{i=1}^{n} q(d_i|x_i)\Big)\Big(\prod_{A \in N} q_{\tau_A}(\Theta_A|x)\Big)$$

where τ_A is the variational hyperparameter indexing distribution $q(\Theta|x)$
We have two options for deriving the update equations. We can

- use the equations we derived for exponential family equations and plug in or
- write out the lower bound explicitly and take the derivative w.r.t. the variational hyperparameter

First, something to remember about the Dirichlet distribution:

- a Dirichlet R.V. $X = [x_1, ..., x_n]$ can be constructed by
- drawing $[y_1, ..., y_k]$ from Gamma$(\alpha_i, 1) = \frac{y_i^{\alpha_i - 1} e^{-y_i}}{\Gamma(\alpha_i)}$
- and setting $x_i = \frac{y_i}{\sum_{j=1}^{K} y_j}$
- it is key to remember that the distribution does in fact have 2 parameters $(\alpha_i, 1)$ and that we are choosing to let the second be 1.

4.1 General note on Variational Bayes

- VB has a fundamental connection to EM algorithm
- EM is a special case of VB
  - gives a point estimate while VB gives you the whole distribution, so VB richer
- Broadly, EM consists of 2 steps
  1. Expectation step: Calculate an objective function
  2. Maximization step: optimize parameters to maximize value of objective function
- VB has analogous steps

  switch!
  1. Expectation step: update the global latent variables
     - global latent variables are shared between different data points
     - for example, in the PCFG case, rule weights are global
     - for common K-means algorithm, analogous global variable would be centroid of clusters
  2. Maximization step: update local latent variables given global ones
     - each data point needs its own local latent variable
     - there are more of these updates
     - in PCFG case, local variables are the parses of each data point (sentence)
     - basically assigning a data point (sentence) to a category (parse)
     - for K-means, analogous local variable would be cluster assignment of each data point

4.2 Updates

Let's first do the method which makes use of our exponential family properties:
Luckily in this model, there aren't too many variational distributions to update. Let's start with the E-step: updating $\tau_{A\to\beta}$, the variational parameter for $q_\tau(\Theta)$. According to our previous equation, we should have 2 parameters, but in this case we only have one (remember we set the second one to 1 by default). So our update becomes:

$$\begin{bmatrix} \tau_{A\to\beta} \\ 1 \end{bmatrix} = \mathbb{E}_q \begin{bmatrix} \alpha_{A\to\beta} + \sum_{i=1}^{N} \sum_{d\in\omega(x_i)} t(x_i, d) \\ 1 \end{bmatrix}$$

Because of our mean field assumption, we can say that $t(x_i, d) = t(x_i)t(d) = f_r(d)q(d|x_i)$

$$\begin{bmatrix} \tau_{A\to\beta} \\ 1 \end{bmatrix} = \mathbb{E}_q \begin{bmatrix} \alpha_{A\to\beta} + \sum_{i=1}^{N} \sum_{d\in\omega(x_i)} q(d|x_i) f_{A\to\beta}(d) \\ 1 \end{bmatrix}$$

$$\begin{aligned} \Rightarrow \tau_{A\to\beta} &= \alpha_{A\to\beta} + \sum_{i=1}^{N} \sum_{d\in\omega(x_i)} q(d|x_i)\mathbb{E}_{\Theta\sim q(\Theta|\tau)}[f_{A\to\beta}(d)] \\ &= \alpha_{A\to\beta} + \sum_{i=1}^{N} \sum_{d\in\omega(x_i)} q(d|x_i)\tilde{f}_{A\to\beta}(d) \end{aligned}$$

where $f_r(d)$ is the number of counts of rule r in derivation tree d, $\tilde{f}_r(d)$ is the expected value of $f_r(d)$, and $\Phi(x_i)$ is the set of derivation trees possible for sentence x_i given the grammar. Now we also need the M-step: updating the variational distribution on parses: $q(d|x_i)$.

$$q(d_j|x_i) = \frac{q(d_j, x_i)}{\sum_{\forall d_j} q(d_j, x_i)} = \frac{\prod_{r\in R} \pi(r)^{f_r(d_j)}}{\sum_{d\in\omega(x_i)} \prod_{r\in R} \pi(r)^{f_r(d)}}$$

where

$$\pi(r) = \exp\left\{ \Psi(\tau_r) - \Psi(\sum_{r'\in R} \tau_{r'}) \right\}$$

is the sum of all ways under the Dirichlet of having generated that rule. Note this is equivalent to:

$$\exp\left\{ \int \Theta p(\Theta|\tau) d\Theta \right\} = \exp\left\{ \mathbb{E}_\Theta[p(\Theta|\tau)] \right\} = \\ \exp\left\{ \Psi(\tau_r) - \Psi(\sum_{r'\in R} \tau_{r'}) \right\}$$

since we're working with a Dirichlet R.V. in log space.
Alternating between these updates ($q(\Theta|X)$ and $q(d|X)$) guarantees convergence on a **local** maximum. Hopefully, this is also the **global** maximum. However, the real world tends to be nonconvex, and in this case we run a high risk of converging on a local max.
Easiest way to get better results: run the whole model multiple times with different initializations. In this case, we still do not guarantee convergence on the global optimum, but we have a higher chance of converging on it or on a better local optimum which might be good enough. Blei et al. (2017)

References

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted).

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Kurihara, K. and Sato, T. (2004). An application of the variational bayesian approach to probabilistic context-free grammars. In *IJCNLP-04 Workshop beyond shallow analyses*.