Bayesian IBM Model 1

Philip Schulz

last modified: April 6, 2017

Abstract

This note describes a Bayesian formulation of IBM model 1 with a Dirichlet prior on the translation parameters. A variational inference algorithm for this model is derived. Basic familiarity with IBM model 1 is assumed.

1 IBM Model 1

IBM model 1 is a joint model of m French words and their corresponding alignment links given an English sentence consisting of $l+1$ words.[1] The English sentence has a 0^{th} position which contains a hypothetical NULL word. Since we aim at a Bayesian formulation of the model, we also condition on a parameter vector θ (which in the case of IBM1 contains only the translation parameters). The joint probability of a French sentence plus alignment is

$$P(f_1^m, a_1^m | e_0, ^l, \theta) = P(m|l) \prod_{j=1}^{m} P(a_j) P(f_j | e_{a_j}, \theta) \tag{1}$$

$$\propto \prod_{j=1}^{m} P(a_j) P(f_j | e_{a_j}, \theta) \ . \tag{2}$$

Observe that line (1) expresses a set of independence assumptions, namely that the French words are conditionally independent given their alignment links and that the alignment links are independent of all other variables. Line (2) further expresses the assumption that the probability for the French sentence length is uniform over some appropriately chosen support.[2] Recall that IBM1 also assumes that $P(a_j)$ is uniform for all $j \in \{1, 2, \ldots, m\}$.

We now rewrite line (2) as a product over English and French word types. This will facilitate the derivation of a variational inference algorithm later

[1] The model and algorithm are presented on the sentence level. Since we standardly assume independence between sentence pairs, the extension to the corpus level is trivial. The names *English* and *French* are used genererically and can be replaced by any other two languages.

[2] This assumption is shared by all IBM models.

on.

$$P(f_1^m, a_1^m | e_0,^l, \theta) \propto P(a_1^m) \prod_e \prod_f P(f_j | e_{a_f})^{\sum_{j=1}^m \mathbb{1}_e(e_{a_j}) \mathbb{1}_f(f_j)} \tag{3}$$

$$= P(a_1^m) \prod_e \prod_f \theta_{f|e}^{\sum_{j=1}^m \mathbb{1}_e(e_{a_j}) \mathbb{1}_f(f_j)} \tag{4}$$

Here we have used the indicator function

$$\mathbb{1}_x(y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise .} \end{cases} \tag{5}$$

Thus, in Equation (4) we simply count the number of times that two word types e and f have been aligned and use that count in the exponent of the conditional categorical distribution associated with e. This gives us the same probability model as Equation (2).

2 A Dirichlet Prior for the Translation Parameters

Since the translation parameters are the only parameters of IBM1, they are also the only ones for which we need to find a prior if we want to build a Bayesian model. The "natural" choice of prior for these categorical parameters is the Dirichlet distribution as it is conjugate to the categorical. Conjugacy means (in this case) that the posterior distribution over parameters will again be a Dirichlet distribution. A K-dimensional Dirichlet distribution is a continous distribution over the probability simplex[3] in $\mathbb{R}^K$ and hence has a probability density function (pdf). Notionally, I distinguish pdfs (p) from probbability mass function (pmfs; P). Thus the type of probability function is always apparent from the formula. The density of the K-dimensional Dirichlet with parameter vector α is

$$p(\theta|\alpha) = \underbrace{\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)}}_{\text{normalisation constant}} \prod_{k=1}^K \theta_k^{\alpha_k - 1} . \tag{6}$$

In the above, $\Gamma(\cdot)$ is the Gamma function, also known as *generalized factorial function*. Denoting the French vocabulary size by V_F, we know that we need a V_F-dimensional Dirichlet prior for our Bayesian IBM1 model. The joint distribution of alignment links, French words and translation parameters given the English sentence and the Dirichlet parameter vector α is

[3] The probability simples is the set of all positive vectors in $\mathbb{R}^K$ with norm 1.

given below. Notice that because the parameters are continuous, this joint distribution is also continuous.

$$p(a_1^m, f_1^m, \theta | e_0^l, \alpha) = P(a_1^m) \prod_e p(\theta_e|\alpha) \prod_f P(f_j|e_{a_f})^{\sum_{j=1}^m \mathbb{1}_e(e_{a_j})\mathbb{1}_f(f_j)} \tag{7}$$

Observe that this Bayesian formulation is not much different from Equation (3), except that we now have distributions over the categorical parameters. Next, we replace the terms $p(\theta_e|\alpha)$ and $P(f_j|e_{a_f})$ with their expansions from Equations (6) and (4). This gives us the functional form of the joint distribution.

$$p(a_1^m, f_1^m, \theta | e_0^l, \alpha) = P(a_1^m) \prod_e \frac{\Gamma\left(\sum_f \alpha_f\right)}{\prod_f \Gamma(\alpha_F)} \prod_f \theta_{f|e}^{\alpha_f - 1} \prod_f \theta_{f|e}^{\sum_{j=1}^m \mathbb{1}_e(e_{a_j})\mathbb{1}_f(f_j)} \tag{8}$$

$$= P(a_1^m) \prod_e \frac{\Gamma\left(\sum_f \alpha_f\right)}{\prod_f \Gamma(\alpha_F)} \prod_f \theta_{f|e}^{\alpha_f - 1 + \sum_{j=1}^m \mathbb{1}_e(e_{a_j})\mathbb{1}_f(f_j)} \tag{9}$$

$$\propto P(a_1^m) \prod_e \prod_f \theta_{f|e}^{\alpha_f - 1 + \sum_{j=1}^m \mathbb{1}_e(e_{a_j})\mathbb{1}_f(f_j)} \tag{10}$$

The expression in line (10) is satisfyingly compact and intuitively easy to understand. The prior parameters effectively act as a priori known alignment links between the English and French types. The reason we were allowed to drop the rather nasty-looking normalisation constant of the Dirichlet distribution is that it is a constant with respect to the probability of the alignment links, French words and parameters.

Equation (10) also has important implications for the computation of the posterior over the categorical parameters. From the definition of conditional probability we know that

$$p(\theta | e_0^l, f_1^m, a_1^m, \alpha) \propto p(\theta, f_1^m, a_1^m | e_0^l, \alpha) \ . \tag{11}$$

Thus the posterior is proportional to line (10). Now observe that the last product in that line has almost the same form as the Dirichlet distribution in Equation (6). The only thing that is missing to make it a proper Dirichlet Distribtion is the normalisation constant. That the Drichlet is unnormalised does not come as a surprise since Equation (11) tells us exactly that we have computed an unnormalized posterior! In order to normalize the posterior we need to fill in the correct normalisation constant which can be done easily if we know the Dirichlet parameter of the posterior. Taking another look at Equation (6), we realise that the f^{th} dimension of the posterior Dirichlet parameter vector is equal to $\alpha_f + \sum_{j=1}^m \mathbb{1}_e(e_{a_j})\mathbb{1}_f(f_j)$. Thus the normalisation constant for the Dirichlet posterior associated with English

type e is

$$\frac{\Gamma\left(\sum_f \alpha_f + \sum_{j=1}^m \mathbb{1}_e(e_{a_j})\mathbb{1}_f(f_j)\right)}{\prod_f \Gamma(\alpha_f + \sum_{j=1}^m \mathbb{1}_e(e_{a_j})\mathbb{1}_f(f_j))} \,. \tag{12}$$

Finally, notice that to get the actual posterior we would need another constant related to the term $P(a_j)$. Since this term is a constant and does not influence posterior inference, we will neglect it though. Before we show how to do posterior inference for the Bayesian IBM model 1, we do a quick recap of exponential family distributions.

3 The Exponential Family

First off, there is not one exponential family but several (one for each distribution that can be written as an exponential family). What's special about exponential family distributions is that their density (up to a constant) is fully determined by their parameters and sufficient statistics. This means that during inference, we only need to collect the sufficient statistics from the data and can ignore all other information that the data may contain. An exponential family distribution with canonical parameter vector θ and sufficient statistics $t(x)$ is any distribution whose density can be written as

$$p(x|\theta) = h(x)\exp\left(\eta(\theta)^\top t(x) - a(\theta)\right) \tag{13}$$

where

- $h(x)$ is a base measure that only depends on the data x
- $t(x)$ is the vector of sufficient statistics
- $\eta(\theta)$ is a vector of *natural* parameters
- $a(\theta) = \log\left(\int h(x)\exp\left(\eta(\theta)^\top t(x)\right) d\theta\right)$ is the log-normaliser that only depends on the parameters.

When we are dealing with pmfs instead of pdfs, the integral is turned into a sum.

The exponential familiy forms of the categorical and Dirichlet distributions are given below.[4] They should be derived as an exercise. Notice that in both cases $h(x) = 1$. We use the count function $c(\cdot)$ which maps a data vector to a vector of counts of unique outcomes.

$$P(x|\theta) = \exp\left(\log(\theta)^\top c(x)\right) \tag{14}$$

[4]Readers familiar with exponential family distributions may remark that the formulations below are not actual exponential families since the categorical parameters are not identifiable. While I acknowledge that they are correct, we ignore the issue of identifiability here as it does not affect the inference algorithm.

$$p(\theta|\alpha) = \exp\left((\alpha-1)^\top \log(\theta) - \left\{\sum_{k=1}^{K} \log\left(\Gamma(\alpha_k)\right) - \log\left(\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\right)\right\}\right) \quad (15)$$

Observe that the sufficient staticistics of the Dirichlet and the natural parameters of the the categorical are identical. Thus, when we multiply both distributions we end up with an unnormalized distribution with sufficient statistics $\log(\theta)$ and natural parameters $\alpha + c(x)$ which is exactly the posterior parameter that we found in Section 2.

A further advantage of exponential family distributions is that their *expected* sufficient statistics are really easy to compute. In particular, the following equivalence holds:

$$\mathbb{E}\left[t(x)\right] = \frac{d}{d\eta(\theta)} a(\eta(\theta)) \quad (16)$$

We do not prove the equivalence here, but if you are able to differentiate Equation (13) you can easily prove it yourself. For the Dirichlet distribution this equivalence lets us compute the expected sufficient statistics as

$$\begin{aligned}
\mathbb{E}\left[\log(\theta_k)\right] &= \frac{\partial}{\partial\alpha_k} a(\alpha) && (17)\\
&= \frac{\partial}{\partial\alpha_k}\sum_{k=1}^{K}\log\left(\Gamma(\alpha_k)\right) - \frac{\partial}{\partial\alpha_k}\log\left(\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\right) && (18)\\
&= \Psi(\alpha_k) - \Psi\left(\sum_{k=1}^{K}\alpha_k\right) && (19)
\end{aligned}$$

where $\Psi(\cdot) = \frac{\Gamma'(\cdot)}{\Gamma(\cdot)}$ is the *digamma function* (which is the first derivative of the log-gamma function).

4 Variational Inference for the Bayesian IBM1

4.1 General Variational Inference

We are now in a position to develop a variational inference algorithm for the Bayesian IBM1. Let us first recall the logic of variational inference: in sufficiently complex models, we cannot compute the posterior because we cannot compute its normalisation constant $p(x) = \int p(x,\theta)d\theta$. Thus, instead of directly computing the posterior $p(\theta|x)$ we decide to approximate it with a simpler distribution $q(\theta|\lambda)$ (which in principle may also depend on x). We measure the divergence of q from p with Kullback-Leibler (KL) divergence (also known as *relative entropy*). Small KL divergence means that q closely approximates p. As a recap, we give the KL formula here.

$$\mathrm{KL}\left(q||p\right) = \int q(\theta|\lambda)\log\left(\frac{q(\theta|\lambda)}{p(\theta|x)}\right)d\theta = \mathbb{E}_q\left[\log\left(\frac{q(\theta|\lambda)}{p(\theta|x)}\right)\right] \quad (20)$$

In order to find a q that is as close as possible to p, we seek to minimise this KL divergence. In practice, however, we maximise the negative KL divergence. This obviously does not affect the q that we find but is motivated by the fact that it allows us to compute a lower bound on the log marginal likelihood of the data Note that the marginal likelihood is also called *evidence*. For this reason, the bound on the evidence is referred to as *evidence lower bound* (ELBO).

$$
\begin{aligned}
-\mathrm{KL}\left(q||p\right) &= \int q(\theta|\lambda) \log\left(\frac{p(\theta|x)}{q(\theta|\lambda)}\right) d\theta && (21)\\
&= \mathbb{E}_q\left[\log(p(\theta, x)) - \log(q(\theta|\lambda))\right] - \log(p(x)) && (22)\\
&= \underbrace{\mathbb{E}_q\left[\log(p(\theta, x))\right] + \mathbb{H}\left(q\right)}_{\text{ELBO}} - \log(p(x)) && (23)
\end{aligned}
$$

From the above it can easily be seen that the ELBO is indeed a lower bound. The gap between the ELBO and the log-evidence is exactly the KL divergence between the approximate posterior $q(\theta|\lambda)$ and the model posterior $p(\theta|x)$. The ELBO contains all the terms from the KL divergence that depend on the variational distribution. Thus it suffices to optimise only the ELBO in order to optimise the KL divergence! The ELBO will from here on serve as our objective.

4.2 The Case of IBM1

In the case of Bayesian IBM1, the situation is a bit more complicated as we have two sets of latent variables, namely the alignment links a_j and the translation parameters θ_e. We thus need two kinds of variational distributions $q(a_j|\phi_j)$ and $q(\theta_e|\lambda_e)$. Since we want to minimize the KL to the model posterior, it seems wise to choose variational distributions that are in the same parametric family as that posterior. Since $P(a_j|e_0^l, f_1^m, \theta, \alpha)$ is categorical and $p(\theta_e|e_0^l, f_1^m, a_1^m, \alpha)$ is Dirichlet, so are their corresponding variational distributions. Notice further that each alignment link and each translation distribution have their own variational distributions. This means that under the variational approximation, all alignment links and all translation distributions are completely independent of other variables. This independence assumption is known as *mean field assumption*. It is a rather harsh assumption, however, it makes inference easy and is therefore widly used. It can be formally expressed as

$$
q(\theta, a_1^m|\phi, \lambda) = q(a_1^m|\phi)q(\theta|\lambda) = \prod_e q(\theta_e|\lambda_e) \prod_{j=1}^{m} q(a_j|\phi_e)\ . \qquad (24)
$$

We can now fomrulate the ELBO for the Bayesian IBM model 1. All expectations are taken with respect to the joint variational distribution from

Equation (24).

$$\begin{aligned}\text{ELBO}(\phi,\lambda) = \sum_{j=1}^{m} \mathbb{E}_q \left[\log(\overbrace{P(a_j)}^{\text{constant}} P(f_j|e_{a_j},\theta)) - q(a_j|\phi) \right] \\ + \sum_e \mathbb{E}_q \left[\log(p(\theta_e|\alpha)) - q(\theta_e|\lambda_e)\right]\end{aligned} \tag{25}$$

Recall that IBM1 assumes $P(a_j)$ to be constant and thus independent of the variational parameters. We can hence drop it during optimisation. In order to optimise (read: maximise) the ELBO, we need to take partial derivatives with respect to each variational parameter and set these derivatives to 0.[5] Let us start with the ϕ parameters. We write $\mathbb{E}_{q_\lambda}[\cdot]$ to denote an expectation only with respect to $q(\theta|\lambda)$.

$$\begin{aligned}\frac{\partial}{\partial\phi_k}\text{ELBO}(\phi,\lambda) = \frac{\partial}{\partial\phi_k}\sum_{j=1}^{m}\mathbb{E}_q\left[\log(P(f_j|e_{a_j},\theta) - \log(q(a_j|\phi))\right] \\ + \frac{\partial}{\partial\phi_k}\sum_e \mathbb{E}_q\left[\log(p(\theta_e|\alpha)) - \log(q(\theta_e|\lambda_e))\right]\end{aligned} \tag{26}$$

$$= \mathbb{E}_{q_\lambda}\left[\log(P(f_k|e_{a_k},\theta)) - \log(q(a_k|\phi_k)) + C\right] \tag{27}$$

$$= \mathbb{E}_{q_\lambda}\left[\log\left(\theta_{f|e_{a_k}}\right)\right] - \log(q(a_k|\phi_k)) + C \tag{28}$$

The constant C absorbs the Dirichlet normaliser (see Equation (6)) and other numeric constant incurred during differentiation.

At this point it is important to realise that the remaining expectation in line (28) is exactly the expected sufficient statistics of the variational Dirichlet distribution. Equation (19) shows us how to compute this term. Setting Equation (28) to 0, we conclude that the update that maximises the ELBO with respect to ϕ_k is

$$q(a_k|\phi_k) \propto \exp\left(\Psi\left(\lambda_{f_k|e_{a_k}}\right) - \Psi\left(\sum_f \lambda_{f|e_{a_k}}\right)\right) . \tag{29}$$

The nessecary normalisation constant can be computed by summing over all possible values for a_k which are the English positions: $\{0,1,\ldots,l\}$. The final update equation is given below.

$$q(a_k|\phi_k) = \frac{\exp\left(\Psi\left(\lambda_{f_k|e_{a_k}}\right) - \Psi\left(\sum_f \lambda_{f|e_{a_k}}\right)\right)}{\sum_{i=0}^{l}\exp\left(\Psi\left(\lambda_{f_k|e_{a_i}}\right) - \Psi\left(\sum_f \lambda_{f|e_{a_i}}\right)\right)} \tag{30}$$

[5] Usually, we would also like to verify the concavity of our objective. However, since we are optimising the KL divergence we are guaranteed to find a local maximum. See Neal and Hinton (1999) for a proof.

4.3 Connection to EM

Notice the parallel to the EM algorithm for the IBM models: there we used the model posterior over alignment links to compute the expected sufficient statistics in the E-step. In variational inference we do exactly the same when optimizing the ϕ parameters. The only difference is that we now compute the expected sufficient statistics under the variational approximation instead of the model posterior.

The analogy extends to the update of the variational Dirichlet distributions $q(\theta|\lambda)$: These are updated in alternation with the update from Equation (30). Since they are distributions over parameters we can interpret their update as an M-step. Let me stress, though, that in Bayesian statistics the distinction between parameters and latent variables does not exist (parameters are just latent variables themselves) and the M-step analogy should not be taken too seriously.

4.4 Optimising the Dirichlet Parameters

In order to compute the updates for the variational Dirichlet parameters, we need to differentiate the ELBO with respect to these parameters, equate the result to 0 and solve for $q(\theta|\lambda)$. This is the same reasoning that we already employed for the updates of $q(a_j|\phi_j)$. We use $\mathbb{E}_{q_\phi}[\cdot]$ to denote expectations taken only with respect to $q(a_1^m|\phi)$. To make our lifes easier, we will also reformulate the model distributions in the ELBO (the p distributions) in the same style as in Section 2. The ELBO is then written as a sum over the English and French vocabularies. The constant for the (uniform) alignment probabilities is ommited as it does not depend on any of the variational parameters.

$$\frac{\partial}{\partial \lambda_k} \text{ELBO}(\phi, \lambda) = \frac{\partial}{\partial \lambda_k} \left\{ \sum_e \mathbb{E}_q \left[p(\theta_e|\alpha_e) - q(\theta_e|\lambda_e) \right] \right. \tag{31}$$
$$\left. + \sum_f \mathbb{E}_q \left[\log(p(f|\theta_e)) \sum_{j=1}^{m} \mathbb{1}_e(e_{a_j}) \mathbb{1}_f(f_j) - \log(q(a_j|\phi_j)) \right] \right\}$$
$$= \frac{\partial}{\partial \lambda_k} \sum_e \sum_f \mathbb{E}_q \left[\log(\theta_{f|e})(\alpha_f - 1) - q(\theta_e|\lambda_e) \right. \tag{32}$$
$$\left. + \log\left(\theta_{f|e}\right) \sum_{j=1}^{m} \mathbb{1}_e(e_{a_j}) \mathbb{1}_f(f_j) - \log(q(a_j|\phi_j)) \right]$$
$$= \frac{\partial}{\partial \lambda_k} \sum_e \sum_f \mathbb{E}_q \left[\log(\theta_{f|e}) \left(\alpha_f - 1 + \sum_{j=1}^{m} \mathbb{1}_e(e_{a_j}) \mathbb{1}_f(f_j) \right) \right. \tag{33}$$
$$\left. - \log(q(\theta_e|\lambda_e)) - \sum_{j=1}^{m} \log(q(a_j|\phi_j)) \right]$$
$$= \sum_f \mathbb{E}_{q_\phi} \left[\log(\theta_{f|k}) \left(\alpha_f - 1 + \sum_{j=1}^{m} \mathbb{1}_k(e_{a_j}) \mathbb{1}_f(f_j) \right) - \log(q(\theta_k|\lambda_k)) + C \right] \tag{34}$$
$$= \sum_f \log(\theta_{f|k}) \left(\alpha_f - 1 + \sum_{j=1}^{m} \mathbb{E}_{q_\phi} \left[\mathbb{1}_k(e_{a_j}) \mathbb{1}_f(f_j) \right] \right) - \log(q(\theta_k|\lambda_k)) + C \tag{35}$$

Observe that we have again exploited conjugacy, just as in Section 2.[6] Through setting the above derivative to 0 and solving for $q(\theta_k|\lambda_k)$ we find that

$$q(\theta_k|\lambda_k) \propto \exp\left(\sum_f \log(\theta_{f|k}) \left(\alpha_f - 1 + \sum_{j=1}^{m} \mathbb{E}_{q_\phi} \left[\mathbb{1}_k(e_{a_j}) \mathbb{1}_f(f_j) \right] \right) \right) \tag{36}$$

which we again recognise as an unnormalised Dirichlet distribution. Since for the Dirichlet the normalisation constant can directly be computed given the vector of natural parameters, this update is often written only in term of those natural parameters.

$$\lambda_{f|k} = \alpha_f + \sum_{j=1}^{m} \mathbb{E}_{q_\phi} \left[\mathbb{1}_k(e_{a_j}) \mathbb{1}_f(f_j) \right] \tag{37}$$

Thus, the variational Dirichlet parameters for English word k are simply the sum of the model's Dirichlet parameters and the expected sufficient statistics

[6] In fact, we have repeated the computation of the Dirichlet posterior inside the computation of the derivative. I hope that this redundancy helps the reader's understanding.

of the conditional categorical distribution associated with k. The expected sufficient statistics are the expected number of times that k is aligned to each French word. The expectation is taken with respect to the variational distribtion over alignment links $q(a_1^m|\phi)$.

5 Summary

We have shown how to compute a posterior Dirichlet distribution for a Dirichlet-Categorical model (IBM1). We have further shown how to derive a variational inference algorithm for that model. While the derivation is somewhat involved, the resulting update equations are pleasingly simple. The implementation of the corresponding algorithm can be done analogously to the EM algorithm (and in fact only requires the modification of a couple of lines of code if the EM algorithm is already in place). For the convenience of the reader, the update equations are repeated below.

$$q(a_k|\phi_k) = \frac{\exp\left(\Psi\left(\lambda_{f_k|e_{a_k}}\right) - \Psi\left(\sum_{f=1}^{V_F}\lambda_{f|e_{a_k}}\right)\right)}{\sum_{i=0}^{l}\exp\left(\Psi\left(\lambda_{f_k|e_{a_i}}\right) - \Psi\left(\sum_{f=1}^{V_F}\lambda_{f|e_{a_i}}\right)\right)} \tag{38}$$

$$\lambda_{f|k} = \alpha_f + \sum_{j=1}^{m}\mathbb{E}_{q_\phi}\left[\mathbb{1}_k(e_{a_j})\mathbb{1}_f(f_j)\right] \tag{39}$$

References

Radford M. Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, 1999.