Maximum Likelihood Estimation for IBM 1 and 2

Wilker Aziz

December 8, 2016

In these notes I do not discuss EM in depth with proofs and guarantees, but I go step by step over the derivation of the M-step for models such as IBM 1 and 2.

1 Notation

Let $\vec{F}$ be a random sequence of m French words representing a random French sentence, and $\vec{f} = \langle f_1 \dots f_m \rangle$ an assignment of this random variable. Similarly, let $\vec{E}$ be a random sequence of English words representing a random English sentence, and $\vec{e} = \langle e_1 \dots e_l \rangle$ an assignment. Finally, let $\vec{A}$ be a random vector over alignments, where an alignment is bijection that maps from $[1 \dots m]$ to $[0 \dots l]$. Note that we have extended every English sentence to contain a NULL token occupying the 0th position.

2 IBM model 1

Equation (1e) specifies IBM model 1(Brown et al., 1993). Assumptions: 1) alignments are independent of one another; 2) the distribution over possible alignments is uniform. The lexical distribution $P(F|E)$ is parameterised as a collection of categorical distributions. That is, let $P(f|e) = \theta_{e,f}$ where $e \in V_E \cup \{\text{NULL}\}$ is a word in the English vocabulary (or NULL) and $f \in V_F$ is a word in the French vocabulary, then $\sum_f \theta_{e,f} = 1$. In Equation (1a), we marginalise over possible alignments. In Equation (1b), we make use of the fact that IBM 1 and 2 generate word-and-alignment pairs independently. In Equation (1c), we assume French words independent of all but the one English word it aligns to. In Equation (1d), we use the independence assumption to rearrange the sum (a straightforward result to readers familiar with mixture models). Up until here IBM models 1 and 2 are identical. In the last step (Equation (1e)), we make the alignment distribution uniform, which is particular of IBM model 1.

$$P(\vec{F} = \vec{f} | \vec{E} = \vec{e}, L = l, M = m) = \sum_{\vec{a}} P(\vec{f}, \vec{a} | \vec{e}, l, m) \tag{1a}$$

$$= \sum_{\vec{a}} \prod_{j=1}^{m} P(f_j, a_j | \vec{e}, l, m) \tag{1b}$$

$$= \sum_{\vec{a}} \prod_{j=1}^{m} P(a_j | l, m) P(f_j | e_{a_j}) \tag{1c}$$

$$= \prod_{j=1}^{m} \sum_{i=0}^{l} P(a_j = i | l, m) P(f_j | e_i) \tag{1d}$$

$$\propto \prod_{j=1}^{m} \sum_{i=0}^{l} P(f_j | e_i) \tag{1e}$$

In order to derive MLE estimates for IBM1, let us pretend for a moment that we observe a single sentence pair.[1] Then Equation (2) states the maximum likelihood objective, a constrained optimisation.

$$\theta_{\text{MLE}} = \arg\max_{\Theta} \prod_{j=1}^{m} \sum_{i=0}^{l} \theta_{e_i, f_j} \tag{2}$$
$$\text{s.t. } \forall e, \sum_{f} \theta_{e,f} = 1$$

We can approach the optimisation problem in Equation (2) as an unconstrained optimisation by introducing a collection of Lagrangian multipliers (one per categorical distribution). Thus, let λ_e for $e \in V_E \cup \{\text{NULL}\}$ be a Lagrangian multiplier.

$$h(\theta, \lambda) = \prod_{j=1}^{m} \sum_{i=0}^{l} \theta_{e_i, f_j} - \sum_{e} \lambda_e \left(\sum_{f} \theta_{e,f} - 1 \right) \tag{3}$$

We can now take derivatives of Equation (3) with respect to some $\theta_{e,f}$ and λ_e and set those to zero.

Let us start with the likelihood term.

[1] This is fine as long as we assume our dataset to be made of iid sentence pairs.

$$\frac{\partial h(\theta,\lambda)}{\partial \theta_{e,f}} = \sum_{\vec{a}} \frac{\partial}{\partial \theta_{e,f}} P(\vec{f},\vec{a}|\vec{e}) \tag{4}$$

$$= \sum_{\vec{a}} P(\vec{f},\vec{a}|\vec{e}) \sum_{j=1}^{m} \frac{\partial}{\partial \theta_{e,f}} \log \theta_{e_{a_j},f_j} \tag{5}$$

$$= \sum_{\vec{a}} P(\vec{f},\vec{a}|\vec{e}) \sum_{j=1}^{m} \frac{1}{\theta_{e,f}} \delta_{\{e\}}(e_{a_j}) \delta_{\{f\}}(f_j) \tag{6}$$

$$= \frac{1}{\theta_{e,f}} \sum_{\vec{a}} P(\vec{f},\vec{a}|\vec{e}) \sum_{j=1}^{m} \delta_{\{e\}}(e_{a_j}) \delta_{\{f\}}(f_j) \tag{7}$$

$$= \frac{1}{\theta_{e,f}} \sum_{\vec{a}} P(\vec{f},\vec{a}|\vec{e}) \eta(e \to f|\vec{a}) \tag{8}$$

$$= \frac{1}{\theta_{e,f}} \sum_{\vec{a}} P(\vec{f}|\vec{e}) P(\vec{a}|\vec{f},\vec{e}) \eta(e \to f|\vec{a}) \tag{9}$$

$$= \frac{P(\vec{f}|\vec{e})}{\theta_{e,f}} \sum_{\vec{a}} P(\vec{a}|\vec{f},\vec{e}) \eta(e \to f|a) \tag{10}$$

$$= \frac{P(\vec{f}|\vec{e})}{\theta_{e,f}} \left\langle \eta(e \to f|\vec{a}) \right\rangle_{P(\vec{a}|\vec{f},\vec{e})} \tag{11}$$

In Equation (4), we push the derivative through the sum due to linearity, but stop at $P(\vec{f},\vec{a}|\vec{e})$ which involves derivatives of products. To deal with the product, in Equation (5) we make use of the log-identity for derivatives, i.e $\frac{\mathrm{d}}{\mathrm{d}x} \log f(x) = \frac{1}{f(x)} \frac{\mathrm{d}}{\mathrm{d}x} f(x)$, thus $\frac{\mathrm{d}}{\mathrm{d}x} f(x) = f(x) \frac{\mathrm{d}}{\mathrm{d}x} \log f(x)$. Again due to linearity, the derivative goes through the sum and stops at the log. In Equation (6), we take the derivative of the log with respect to $\theta_{e,f}$, which will be non-zero only when e_{a_j} matches the context e and f_j matches the decision f —to express this fact we introduce $\delta_A(a)$ which is 1 when $a \in A$ and 0 otherwise. In Equation (7) we just rearrange the terms making it explicit that $\theta_{e,f}$ does not depend on $\vec{a}$ or j. In Equation (8) we introduce a function $\eta(e \to f|\vec{a}) = \sum_{j=1}^{m} \delta_{\{e\}}(e_{a_j}) \delta_{\{f\}}(f_j)$ which counts the number of times e generates f in $\vec{a}$. Equation (9) follows by application of the chain rule of probabilities. And Equation (10) follows because the marginal $P(\vec{f}|\vec{e})$ is not a function of $\vec{a}$. This last result is really handy as it leaves us with a sum over $P(\vec{a}|\vec{f},\vec{e})\eta(e \to f|\vec{a})$ where $P(\vec{a}|\vec{f},\vec{e})$ is the posterior probability over alignments and the sum is in fact an expectation. Equation (11) shows the most important result of this block of identities, namely, that the derivative is proportional to the expected number of occurrences of e, f under the posterior distribution over alignment configurations.

Now let us turn to the Lagrangian term. Its derivative with respect to a fixed $\theta_{e,f}$ is shown in Equation (15).

$$\frac{\partial}{\partial \theta_{e,f}} h(\theta, \lambda) = \frac{\partial}{\partial \theta_{e,f}} \sum_{e'} \lambda_{e'} \sum_{f'} \theta_{e',f'} - 1 \tag{12}$$

$$= \sum_{e'} \lambda_{e'} \sum_{f'} \frac{\partial}{\partial \theta_{e,f}} \theta_{e',f'} - 0 \tag{13}$$

$$= \sum_{e'} \lambda_{e'} \sum_{f'} \delta_{\{e\}}(e') \delta_{\{f\}}(f') \tag{14}$$

$$= \lambda_e \tag{15}$$

Now we put together Equation (11) (derivative of the likelihood term) and Equation (15) (derivative of the Lagrangian term), and make $\frac{\partial}{\partial \theta_{e,f}} h(\theta, \lambda) = 0$.

$$0 = \frac{\partial}{\partial \theta_{e,f}} h(\theta, \lambda) \tag{16}$$

$$0 = \frac{P(\vec{f}|\vec{e})}{\theta_{e,f}} \langle \eta(e \to f|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})} - \lambda_e \tag{17}$$

$$\lambda_e = \frac{P(\vec{f}|\vec{e})}{\theta_{e,f}} \langle \eta(e \to f|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})} \tag{18}$$

$$\theta_{e,f} = \frac{P(\vec{f}|\vec{e})}{\lambda_e} \langle \eta(e \to f|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})} \tag{19}$$

Now recall the constraint $\sum_f \theta_{e,f} = 1$ which can also be derived by taking $\frac{\partial}{\partial \lambda_e} h(\theta, \lambda) = 0$.

$$1 = \sum_f \theta_{e,f} \tag{20}$$

$$= \sum_f \frac{P(\vec{f}|\vec{e})}{\lambda_e} \langle \eta(e \to f|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})} \tag{21}$$

$$= \frac{P(\vec{f}|\vec{e})}{\lambda_e} \sum_f \langle \eta(e \to f|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})} \tag{22}$$

$$= \frac{P(\vec{f}|\vec{e})}{\lambda_e} \left\langle \sum_f \eta(e \to f|\vec{a}) \right\rangle_{P(\vec{a}|\vec{f},\vec{e})} \tag{23}$$

$$\lambda_e = P(\vec{f}|\vec{e}) \langle \eta(e|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})} \tag{24}$$

In Equation (21), we substitute $\theta_{e,f}$ by the result in Equation (19). We factor $\frac{P(\vec{f}|\vec{e})}{\lambda_e}$ out of the sum since it does not depend on f. Then, we push the sum through the expectation due to linearity. Finally, we define $\eta(e|\vec{a}) \triangleq \sum_f \eta(e \to f|\vec{a})$ as the number of times e is selected in a. The result in Equation (24) can be combined with Equation (19) to yield the final result shown in Equation (25) (note that the marginal $P(\vec{f}|\vec{e})$ cancels out).

$$\theta_{e,f} = \frac{\langle \eta(e \to f|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})}}{\langle \eta(e|\vec{a}) \rangle_{P(\vec{a}|\vec{f},\vec{e})}} \tag{25}$$

Obviously this is not a closed form solution as $\theta_{e,f}$ appears on both sides of the equation —recall that $P(\vec{a}|\vec{f},\vec{e})$ depends parameters θ. But this suggests an iterative solution which is in fact the EM algorithm (Dempster et al., 1977).

References

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.