Maximum Likelihood Estimation for IBM 1 and 2

Wilker Aziz

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 F be a random variable over French sentences and $f = f_1^m = \langle f_1 \dots f_m \rangle$ an assignment of this random variable. Similarly, let E be a random variable over English sentences, and $e = e_1^l$ and assignment. Finally, let A be a random variable over alignments, where an alignment is bijection that maps from $[1 \dots m]$ to $[0 \dots l]$ where we extend every English sentence to contain a Null token occupying the 0th position.

2 IBM model 1

Equation 1 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 t is parameterised as a collection of categorical distributions. That is, let $t(d|c) = \theta_{c,d}$ where $c \in V_E \cup \{\textsc{Null}\}$ is a word in the English vocabulary (or Null) and $d \in V_F$ is a word in the French vocabulary, then $\sum_d \theta_{d,c} = 1$.[1]

$$\begin{aligned} P(F = f|E = e) &= \sum_a P(f, a|e) \\ &\propto \sum_a \prod_{j=1}^{m} t(f_j|e_{a_j}) \end{aligned} \tag{1}$$

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

[1] We use c and d as in *context* and *decision*, respectively.

$$\theta_{\text{MLE}} = \arg\max_{\Theta} \sum_a \prod_{j=1}^m t(f_j|e_{a_j}) \tag{2}$$
$$\text{s.t. } \forall c, \sum_d \theta_{c,d} = 1$$

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_c for $c \in V_E \cup \{\textsc{Null}\}$ be a Lagrangian multiplier.

$$h(\theta, \lambda) = \sum_a \prod_{j=1}^m t(f_j|e_{a_j}) - \sum_c \lambda_c \left(\sum_d \theta_{c,d} - 1 \right) \tag{3}$$

We can now take derivatives of Equation 3 with respect to some $\theta_{c,d}$ and λ_c and set those to zero.

Let us start with the likelihood term.

$$
\begin{aligned}
\frac{\partial h(\theta, \lambda)}{\partial \theta_{c,d}} &= \sum_a \frac{\partial}{\partial \theta_{c,d}} P(f, a|e) & (4)\\
&= \sum_a P(f, a|e) \sum_{j=1}^m \frac{\partial}{\partial \theta_{c,d}} \log t(f_j|e_{a_j}) & (5)\\
&= \sum_a P(f, a|e) \sum_{j=1}^m \frac{1}{\theta_{c,d}} \delta(c, e_{a_j}) \delta(d, f_j) & (6)\\
&= \frac{1}{\theta_{c,d}} \sum_a P(f, a|e) \sum_{j=1}^m \delta(c, e_{a_j}) \delta(d, f_j) & (7)\\
&= \frac{1}{\theta_{c,d}} \sum_a P(f, a|e) n_a(c, d) & (8)\\
&= \frac{1}{\theta_{c,d}} \sum_a P(f|e) P(a|f, e) n_a(c, d) & (9)\\
&= \frac{P(f|e)}{\theta_{c,d}} \sum_a P(a|f, e) n_a(c, d) & (10)\\
&= \frac{P(f|e)}{\theta_{c,d}} \langle n_a(c, d) \rangle_{P(a|f,e)} & (11)
\end{aligned}
$$

In Equation 4, we push the derivative through the sum due to linearity, but stop at $P(f, a|e)$ 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 $\theta_{c,d}$, which will be non-zero only when e_{a_j} matches the context c and f_j matches the decision d – to express this fact we introduce $\delta(a, b)$ which is 1 when $a = b$ and 0 otherwise. In Equation 7 we just rearrange the terms making it explicit that $\theta_{c,d}$ does not depend on a or j. In Equation 8 we introduce a function $n_a(c, d) = \sum_{j=1}^{m} \delta(c, e_{a_j})\delta(d, f_j)$ which counts the number of times c, d participates in a. Equation 9 follows by application of the chain rule of probabilities. And Equation 10 shows that $P(f|e)$ is not a function of a, in fact, it results from marginalisation over all possible alignment configurations. This last result is really handy as it leaves us with a sum over $P(a|f, e)n_a(c, d)$ where $P(a|f, e)$ 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 c, d under the posterior distribution over alignment configurations.

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

$$
\begin{aligned}
\frac{\partial}{\partial \theta_{c,d}} h(\theta, \lambda) &= \frac{\partial}{\partial \theta_{c,d}} \sum_{c'} \lambda_{c'} \sum_{d'} t(d'|c') - 1 && (12)\\
&= \sum_{c'} \lambda_{c'} \sum_{d'} \frac{\partial}{\partial \theta_{c,d}} t(d'|c') - 0 && (13)\\
&= \sum_{c'} \lambda_{c'} \sum_{d'} \delta(c, c')\delta(d, d') && (14)\\
&= \lambda_c && (15)
\end{aligned}
$$

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

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta_{c,d}} h(\theta, \lambda) && (16)\\
0 &= \frac{P(f|e)}{\theta_{c,d}} \langle n_a(c, d) \rangle_{P(a|f,e)} - \lambda_c && (17)\\
\lambda_c &= \frac{P(f|e)}{\theta_{c,d}} \langle n_a(c, d) \rangle_{P(a|f,e)} && (18)\\
\theta_{c,d} &= \frac{P(f|e)}{\lambda_c} \langle n_a(c, d) \rangle_{P(a|f,e)} && (19)
\end{aligned}
$$

Now recall the constraint $\sum_d \theta_{c,d} = 1$ which can also be derived by taking $\frac{\partial}{\partial \lambda_c} h(\theta, \lambda) = 0$.

$$
\begin{aligned}
1 &= \sum_d \theta_{c,d} && (20)\\
&= \sum_d \frac{P(f|e)}{\lambda_c} \langle n_a(c,d) \rangle_{P(a|f,e)} && (21)\\
&= \frac{P(f|e)}{\lambda_c} \sum_d \langle n_a(c,d) \rangle_{P(a|f,e)} && (22)\\
&= \frac{P(f|e)}{\lambda_c} \left\langle \sum_d n_a(c,d) \right\rangle_{P(a|f,e)} && (23)\\
\lambda_c &= P(f|e) \langle n_a(c) \rangle_{P(a|f,e)} && (24)
\end{aligned}
$$

In Equation 21, we substitute $\theta_{c,d}$ by the result in Equation 19. We factor $\frac{P(f|e)}{\lambda_c}$ out of the sum since it does not depend on d. Then, we push the sum through the expectation due to linearity. Finally, we define $n_a(c) \triangleq \sum_d n_a(c,d)$ as the number of times c participates 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 terms $P(f|e)$ cancel out).

$$
\theta_{c,d} = \frac{\langle n_a(c,d) \rangle_{P(a|f,e)}}{\langle n_a(c) \rangle_{P(a|f,e)}} \tag{25}
$$

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

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.