Derivation of the mapping optimization algorithm in the paper

March 29, 2012

Recall that our objective is given by (we neglect the $F_{verb}(A)$ since it is convex and adding it is straightforward):

$$F(A) = D_{KL}[p_S(c_1, c_2)|p_T(c_1, c_2; A)] + \min_{r(c_1,c_2)\in\mathcal{S}} D_{KL}[r(c_1, c_2)|p_T(c_1, c_2; A)] + H[A] \tag{1}$$

We assume the weights α_i and the coefficient λ are one for simplicity. Other values can easily be plugged in.

Since we are minimizing over A we can recast the problem as minimization over both A and $r \in \mathcal{S}$ of the objective:

$$F(A, r) = D_{KL}[p_S(c_1, c_2)|p_T(c_1, c_2; A)] + D_{KL}[r(c_1, c_2)|p_T(c_1, c_2; A)] + H[A] \tag{2}$$

We now recall the variational property of entropy, namely:[1]

$$H[p] = -\sum_x p(x) \log p(x) = \min_q -\sum_x p(x) \log q(x) \tag{3}$$

Where optimization is over distributions q. Thus we can again expand F to contain another variable $q(c|f)$ such that:

$$F(A, r, q) = D_{KL}[p_S(c_1, c_2)|p_T(c_1, c_2; A)] + D_{KL}[r(c_1, c_2)|p_T(c_1, c_2; A)] - \sum_{f,c} A(c|f) \log q(c|f) \tag{4}$$

Clearly $F(A, r, q) \geq F(A)$ for all r, q and $\min_{r,q} F(A, r, q) = F(A)$. Thus, we can proceed in alternating optimization over A, r, q.

We can now see how the algorithm in the paper is obtained. Denote by A^k, r^{k-1}, q^{k-1} the values of these variables at iteration k. Then:

$$r^k(c_1, c_2) = \arg \min_{r(c_1,c_2)\in\mathcal{S}} D_{KL}[r(c_1, c_2)|p_T(c_1, c_2; A^k)] \tag{5}$$

This clearly corresponds to steps 1 and 2 in the algorithm.

[1] This is a direct result of $D_{KL}[p|q] \geq 0$ and zero if and only if $p = q$.

Next, we optimize over q, which results in:

$$q^k(c|f) = A^k(c|f) \tag{6}$$

for all c, f. We note turn to optimizing over A. The objective as a function of A, given the current q^k, r^k is (up to additive constants)

$$F^k(A) = -\sum_{c_1,c_2} \Big[p_S(c_1,c_2) + r^k(c_1,c_2)\Big] \log p_T(c_1,c_2;A) - \sum_{f,c} A(c|f) \log A^k(c|f)$$

This is non-convex due to the bilinear form of $p_T(c_1,c_2;A)$. To simplify things further we use the standard EM trick and define an auxiliary function:

$$\begin{aligned}\bar{F}^k(A) \quad \equiv \quad & -\sum_{c_1,c_2,f_1,f_2} p(f_1,f_2|c_1,c_2;A^k) \Big[p_S(c_1,c_2) + r^k(c_1,c_2)\Big] \log p_T(c_1,f_1,c_2,f_2;A) \\ & -\sum_{f,c} A(c|f) \log A^k(c|f) + g(A^k)\end{aligned}$$

where $p(f_1,f_2|c_1,c_2;A^k)$ is the posterior calculated in step 3 of the algorithm and $g(A^k)$ is a function of A^k and not A.[2] As in standard EM, it can be shown that $F^k(A) \leq \bar{F}^k(A)$ with equality if $A = A^k$. Thus we can minimize $\bar{F}^k(A)$ over A and decrease the objective $F(A,r,q)$.

Using the notation in step 4 of the paper, this simplifies to:

$$\begin{aligned}\bar{F}^k(A) \quad = \quad & -\sum_{c_1,c_2,f_1,f_2} N^k(c_1,c_2,f_1,f_2) \log p_T(c_1,f_1,c_2,f_2;A) \\ & -\sum_{f,c} A(c|f) \log A^k(c|f) + g(A^k)\end{aligned}$$

We can now use the fact that $p_T(c_1,f_1,c_2,f_2;A)$ factors according to:

$$p_T(c_1,f_1,c_2,f_2;A) = A(c_1|f_1)A(c_2|f_2)p_T(f_1,f_2) \tag{7}$$

to obtain (up to additive constants):

$$\begin{aligned}\bar{F}^k(A) \quad \equiv \quad & -\sum_{c,f} N_1^k(c,f) \log A(c|f) - \sum_{c,f} N_2^k(c,f) \log A(c|f) \\ & -\sum_{f,c} A(c|f) \log A^k(c|f)\end{aligned}$$

And using the definition of M^k in step 5 of the paper, we obtain that:

$$\bar{F}^k(A) \quad \equiv \quad -\sum_{c,f} \Big[M^k(c,f) \log A(c|f) + A(c|f) \log A^k(c|f)\Big]$$

[2] It is given by $g(A^k) = -\sum_{c_1,c_2} \left[p_S(c_1,c_2) + r^k(c_1,c_2)\right] H[p(f_1,f_2|c_1,c_2;A^k)]$.

We now just need to minimize it over A, and this indeed corresponds to step 6 in the algorithm (except for the term $F_{verb}(A)$ which is straightforward to add).

The above establishes that the F objective decreases monotonically with each update. Convergence to local optima can be established as in EM.