Appendix D: Proofs for Chapter 5

Lemma 5.1:

(A) At any iterative cycle of the ID algorithm (Figure 5.1) with $t > 0$, update step ID1 decreases the value of F^{ID} (Eq. 5.1) by

$$\Delta F^{ID1}{}_t \;=\; \textstyle\sum_x p(x) KL[p_{t-1}(c|x) \| p_t(c|x)] \,. \tag{D.1}$$

(B) (following Gilad-Bachrach, Navot & Tishby, 2003) At the iterative cycle of the ID algorithm with any t, update step ID2 decreases the value of F^{ID} by

$$\Delta F^{ID2}{}_t \;=\; \beta \textstyle\sum_c p_t(c) KL[p_t(y|c) \| p_{t-1}(y|c)] \,. \tag{D.2}$$

Proof:

(A) Taking log of both sides of the equality sign in step ID1 (note that $p_t(c|x)$ is never equal to 0), we have:

$$\log p_t(c|x) \;=\; \beta \textstyle\sum_y p(y|x) \log p_{t-1}(y|c) - \log z_t(x,\beta) \,, \tag{D.3}$$

where $z_t(x,\beta)$ is a normalization function, over all values c of C, of terms depending on x, c and β. Extracting $\log z_t(x,\beta)$ from the above equality:

$$\log z_t(x,\beta) \;=\; \beta \textstyle\sum_y p(y|x) \log p_{t-1}(y|c) - \log p_t(c|x) \,. \tag{D.4}$$

Note that although a particular value c is being used in Eq. D.4 (which is in fact true for every c), the actual value of $\log z_t(x,\beta)$ does not depend on any particular value of C.

After performing update step ID1 at time t, which results in the replacement of each $p_{t-1}(c|x)$ with $p_t(c|x)$, the value of F^{ID} changes from $F^{ID}{}_{t-1}$, where all $p(c|x)$ and $p(y|c)$ indexed by $t-1$, to

$$F^{ID}{}_{t-} \;\equiv\; \textstyle\sum_x p(x) \sum_c p_t(c|x) \log p_t(c|x) \;-\; \beta \sum_x p(x) \sum_c p_t(c|x) \sum_y p(y|x) \log p_{t-1}(y|c) \,. \tag{D.5}$$

The value we are interested in, $\Delta F^{ID1}{}_t$, is the difference between $F^{ID}{}_{t-1}$ and $F^{ID}{}_{t-}$:

$$\Delta F^{ID1}{}_t \;=\; F^{ID}{}_{t-1} - F^{ID}{}_{t-} \;\overset{(a)}{=} \tag{D.6}$$

$$\textstyle\sum_x p(x) \sum_c p_{t-1}(c|x) \log p_{t-1}(c|x) \;-\; \beta \sum_x p(x) \sum_c p_{t-1}(c|x) \sum_y p(y|x) \log p_{t-1}(y|c) \;+$$

$$- \textstyle\sum_x p(x) \sum_c p_t(c|x) \left(\beta \sum_y p(y|x) \log p_{t-1}(y|c) - \log z_t(x,\beta) \right)$$

$$+ \;\beta \textstyle\sum_x p(x) \sum_c p_t(c|x) \sum_y p(y|x) \log p_{t-1}(y|c) \;\overset{(b)}{=}$$

$$\textstyle\sum_x p(x) \sum_c p_{t-1}(c|x) \log p_{t-1}(c|x) \;-\; \beta \sum_x p(x) \sum_c p_{t-1}(c|x) \sum_y p(y|x) \log p_{t-1}(y|c) \;+$$

$$+ \textstyle\sum_x p(x) \sum_c p_{t-1}(c|x) \left(\log z_t(x,\beta) \right) \;\overset{(c)}{=}$$

$$\sum_x p(x) \sum_c p_{t-1}(c|x) \log p_{t-1}(c|x) - \beta \sum_x p(x) \sum_c p_{t-1}(c|x) \sum_y p(y|x) \log p_{t-1}(y|c) +$$
$$+ \sum_x p(x) \sum_c p_{t-1}(c|x) \left(\beta \sum_y p(y|x) \log p_{t-1}(y|c) - \log p_t(c|x) \right) =^{(d)}$$
$$\sum_x p(x) KL[p_{t-1}(c|x) \| p_t(c|x)]$$

In the equality chain of Eq. D.6, (a) incorporates Eq. D.3, (b) omits identical terms with opposite signs and replaces, for each x separately, expectation over $p_t(c|x)$ with the identical expectation over $p_{t-1}(c|x)$ (as $\log z_t(x,\beta)$ is independent in C), (c) incorporates Eq. D.4 and (d) again omits opposite sign terms and resorts to the definition of *KL* divergence.

(B) The value we are interested in, $\Delta F^{ID2}{}_t$, is the difference between $F^{ID}{}_{t-}$ (Eq. D.5) and $F^{ID}{}_t$ (where all $p(c|x)$ and $p(y|c)$ are indexed with t):

$$\Delta F^{ID2}{}_t = F^{ID}{}_{t-} - F^{ID}{}_t =^{(a)} \tag{D.7}$$
$$-\beta \sum_x p(x) \sum_c p_t(c|x) \sum_y p(y|x) \log p_{t-1}(y|c) + \beta \sum_x p(x) \sum_c p_t(c|x) \sum_y p(y|x) \log p_t(y|c) =^{(b)}$$
$$\beta \sum_{c,y} \left(\log p_t(y|c) - \log p_{t-1}(y|c) \right) \sum_x p(x) p_t(c|x) p(y|x) =^{(c)}$$
$$\beta \sum_{c,y} \left(\log p_t(y|c) - \log p_{t-1}(y|c) \right) p_t(c,y) =^{(d)}$$
$$\beta \sum_c p_t(c) KL[p_t(y|c) \| p_{t-1}(y|c)]$$

In the equality chain of Eq. D.7, (a) drops the term $\sum_x p(x) \sum_c p_t(c|x) \log p_t(c|x)$ with opposite signs from both $F^{ID}{}_{t-}$ and $F^{ID}{}_t$, (b) just re-orders the terms, (c) uses the conditional independence assumption (Eq. 5.4), which step ID2 happens to maintain, and (d) resorts to the definition of (conditioned) *KL* divergence. □

Lemma 5.2: Stable points of the ID algorithm (i.e. probability distributions that remain unchanged under the update steps: $p_{t+1}(c|x) = p_t(c|x)$ and $p_{t+1}(y|c) = p_t(y|c)$ for all c, x and y) are locally extremal points of F^{ID} (Eq. 5.1).

Proof: Update step ID1 of the ID algorithm can be derived, using the method of *Lagrange multipliers*, as follows:

(1) Convert F^{ID} (Eq. 5.1) to a *Lagrangian* L^{ID1}, by adding to it a Lagrange multiplier λ_x for each x, in order to restrict each probability $p(c|x)$ distribution to sum up to 1: $L^{ID1} = F^{ID} + \sum_x \lambda_x (1 - \sum_c p(c|x))$.

(2) Take derivatives from L^{ID1} with respect to each $p(c|x)$.

(3) Equate each of the resulting terms to 0, extract $p(c|x)$ and set λ_x so that all distributions sum up to 1, to obtain the equation specifying ID1.

The details of this derivation closely resemble the derivation of step IB1 for the IB algorithm (Fig. 5.2), which has been given in several previous works (e.g., Tishby, Pereira & Bialek 1999). The above holds as well with regard to the derivation of update step ID2, substituting $p(c|x)$ by $p(y|c)$, λ_x by λ_c and L^{ID1} by $L^{ID2} = F^{ID} + \sum_c \lambda_c (1 - \sum_y p(y|c))$:

(1) The Lagrangian introducing to F^{ID} the constraint of $p(y|c)$ to sum up to 1 is:

$$L^{ID2} \equiv \sum_x p(x) \sum_c p(c|x) \,(\, \log p(c|x) - \beta \sum_y p(y|x) \log p(y|c) \,) + \sum_c \lambda_c \,(\, 1 - \sum_y p(y|c) \,). \tag{D.8}$$

(2) From L^{ID2}, take derivatives relatively to $p(y|c)$:

$$\frac{\delta L^{ID2}}{\delta p(y\,|\,c)} = -\beta \sum_x p(x)\, p(c|x)\, p(y|x) \,(\, 1 / p(y|c) \,) + \lambda_c\,. \tag{D.9}$$

(3) Equating the above term to 0 and setting $\lambda_c = \beta \sum_x p(x)\; p(c|x) = \beta p(c)$ so that the constraint of $p(y|c)$ to sum up to 1 holds (note that $p(c)$ have here the mere role of a normalization factor), we get the equation underlying step ID2 of the ID algorithm:

$$p(y|c) = (\, 1 / p(c) \,) \sum_x p(x)\; p(c|x)\, p(y|x)\,. \tag{D.10}$$

From the above follows that stable probability distributions $p_t(c|x)$ (i.e., ones satisfying $p_{t+1}(c|x) = p_t(c|x)$) specify extremal value of F^{ID} relatively to fixed $p_t(y|c)$ and vice versa: stable probability distributions $p_t(y|c)$ (satisfying $p_{t+1}(y|c) = p_t(y|c)$) specify extremal value of F^{ID} relatively to fixed $p_t(c|x)$. As the $p_t(c|x)$ and $p_t(y|c)$ are all the parameters and they are all fixed in a stable point, together they form an extremal point of F^{ID}. □

Lemma 5.7: In the update cycle of time t, the four CP algorithm steps CP1, CP2, CP*1, CP*2, decrease the value of F^{CP1} (Eq. 5.17), F^{CP2} (Eq. 5.19), $F^{\mathrm{CP*1}}$ (Eq. 5.24), $F^{\mathrm{CP*2}}$ (Eq. 5.27) by

$$\begin{aligned}
\Delta F^{CP1}{}_t &= \sum_x p(x) KL[p_{t-1}(c|x) \| p_t(c|x)]\,, \\
\Delta F^{CP2}{}_t &= \sum_x p_t(c,w) KL[p_t(y|c,w) \| p_{t-1}(y|c,w)]\,, \\
\Delta F^{\mathrm{CP*1}}{}_t &= \sum_y p(y) KL[p^*{}_{t-1}(c|y) \| p^*{}_t(c|y)]\,, \\
\Delta F^{\mathrm{CP*2}}{}_t &= \sum_y p^*{}_t(c) KL[p^*{}_t(y|c) \| p^*{}_{t-1}(y|c)]\,,
\end{aligned} \tag{D.11}$$

respectively.

Proof: We exemplify the proof by proving the claim with regard to $F^{\mathrm{CP*1}}$ (note the similarity to the proof of lemma 5.1 (A)).

Taking log of both sides of the step CP*1 equality sign, we have:

$$\log p^*_t(c|y) \;=\; \eta \textstyle\sum_w p(w) \log p_{t-1}(y|c,w) - \log z^*_t(y,\eta)\,, \tag{D.12}$$

where $z^*_t(y,\eta)$ is a normalization function over all values c of C of terms depending on y, c and η. We then extract $\log z^*_t(y,\eta)$ from the above equality:

$$\log z^*_t(y,\eta) \;=\; \eta \textstyle\sum_w p(w) \log p_{t-1}(y|c,w) - \log p^*_t(c|y)\,. \tag{D.13}$$

Note that although a particular value c is being used in Eq. D.13 (in fact, it is true for every c), the actual value of $\log z^*_t(y,\eta)$ does not depend on any particular value of C.

After performing update step CP*1 at time t, which results in the replacement of each $p^*_{t-1}(c|y)$ with $p^*_t(c|y)$, the value of $F^{\mathrm{CP*1}}$ changes from $F^{\mathrm{CP*1}}_{t-1}$, where all $p^*(c|y)$ and $p(y|c,w)$ are indexed by $t-1$, to

$$F^{\mathrm{CP*1}}_{t-} \;\equiv\; \textstyle\sum_y p(y) \sum_c p^*_t(c|y) \log p^*_t(c|y) \;-\; \eta \sum_w p(w) \sum_y p(y) \sum_c p^*_t(c|y) \log p_{t-1}(y|c,w)\,. \tag{D.14}$$

The value we are interested in, $\Delta F^{\mathrm{CP*1}}_t$, is the difference between $F^{\mathrm{CP*1}}_{t-1}$ and $F^{\mathrm{CP*1}}_{t-}$:

$$\Delta F^{\mathrm{CP*1}}_t \;=\; F^{\mathrm{CP*1}}_{t-1} - F^{\mathrm{CP*1}}_{t-} \;=^{(a)} \tag{D.15}$$

$$\textstyle\sum_y p(y) \sum_c p^*_{t-1}(c|y) \log p^*_{t-1}(c|y) \;-\; \eta \sum_w p(w) \sum_y p(y) \sum_c p^*_{t-1}(c|y) \log p_{t-1}(y|c,w) \;+$$
$$- \textstyle\sum_y p(y) \sum_c p^*_t(c|y)\,(\;\eta \sum_y p(y|x) \log\; p_{t-1}(y|c,w) - \log z^*_t(y,\eta)\;)$$
$$+\;\; \eta \textstyle\sum_w p(w) \sum_y p(y) \sum_c p^*_t(c|y)\, p_{t-1}(y|c,w) \;\;=^{(b)}$$

$$\textstyle\sum_y p(y) \sum_c p^*_{t-1}(c|y) \log p_{t-1}(c|x) \;-\; \eta \sum_w p(w) \sum_y p(y) \sum_c \; p^*_{t-1}(c|y) \log\; p_{t-1}(y|c,w) \;+$$
$$+ \textstyle\sum_y p(y) \sum_c p^*_{t-1}(c|y)\,(\;\log z^*_t(y,\eta)\;) \;\;=^{(c)}$$

$$\textstyle\sum_y p(y) \sum_c p^*_{t-1}(c|y) \log p^*_{t-1}(c|y) \;-\; \eta \sum_w p(w) \sum_y p(y) \sum_c p^*_{t-1}(c|y) \log\; p_{t-1}(y|c,w) \;+$$
$$+ \textstyle\sum_y p(y) \sum_c p^*_{t-1}(c|y)\,(\;\eta \sum_w p(w) \log p_{t-1}(y|c,w) - \log\; p^*_t(c|y)\;) \;\;=^{(d)}$$

$$\textstyle\sum_y p(y)\, KL[\; p^*_{t-1}(c|y) \,\|\, p^*_t(c|y)\;]$$

In the equality chain of Eq. D.15, (a) incorporates Eq. D.12, (b) drops identical terms with opposite signs and replaces, for each y separately, expectation over $p^*_t(c|y)$ with the identical expectation over $p^*_{t-1}(c|y)$ (as $\log z^*_t(y,\eta)$ is independent of C), (c) incorporates Eq. D.13 and (d) again drops opposite sign terms and resorts to the definition of *KL* divergence.

The proof for the claim regarding update step CP1 is in close correspondence to the proof of lemma 5.1 (A). The proofs for the claims regarding update step CP2 and CP*2 are similar to the proof of lemma 5.1 (B). □

Lemma 5.8: A set of probability distributions that form a stable point of the CP algorithm (i.e., ones that satisfy $p_{t+1}(c|x) = p_t(c|x)$, $p_{t+1}(y|c,w) = p_t(y|c,w)$, $p^*_{t+1}(c|y) = p^*_t(c|y)$ and $p^*_{t+1}(y|c) = p^*_t(y|c)$, for all c, x, y and w) specifies locally extremal points for F^{CP1} with respect to $p(c|x)$ ($p^*(y|c)$ held fixed), F^{CP2} with respect to $p(y|c,w)$ ($p(c|x)$ held fixed), $F^{\mathrm{CP}*1}$ with respect to $p^*(c|y)$ ($p(y|c,w)$ held fixed) and $F^{\mathrm{CP}*2}$ with respect $p^*(y|c)$ ($p^*(c|y)$ held fixed).

Proof: As a demonstrative example, we show that distributions $p^*(c|y)$ that are part of a stable point of the CP algorithm specify locally extremal points for $F^{\mathrm{CP}*1}$ (while $p(y|c,w)$ held fixed). The other parts are similar (see also the proof of Lemma 5.2 above).

We first write explicitly $L^{\mathrm{CP}*1}$, the Lagrangian introducing to $F^{\mathrm{CP}*1}$ for every y the constraint of $p^*(c|y)$ to sum up to 1:

$$L^{\mathrm{CP}*1} \equiv \tag{D.16}$$

$$\textstyle\sum_y p(y) \sum_c p^*(c|y) \left(\log p^*(c|y) - \eta \sum_w p(w) \log p(y|c,w) \right) + \sum_y \lambda^*_y \left(1 - \sum_c p^*(c|y) \right).$$

From $L^{\mathrm{CP}*1}$, we take derivatives relatively to $p^*(c|y)$, considering $p(y|c,w)$ as a constant:

$$\frac{\delta L^{CP*1}}{\delta p^*(c \mid y)} = p(y) \left(\log p^*(c|y) - \eta \textstyle\sum_w p(w) \log p(y|c,w) \right) + p(y)\ p^*(c|y) \left(1 / p^*(c|y) \right) + \lambda^*_y\ . \tag{D.17}$$

Equating the above term to 0 and setting $\lambda^*_y = p(y) \left(\log z^*(y,\eta) - 1 \right)$, so that the constraint of $p^*(c|y)$ to sum up to 1 holds, with a normalization factor $z^*(y,\eta) = \sum_{c'} \prod_w p(y|c',w)^{\eta\, p(w)}$, we get the equation underlying step IB2 of the IB algorithm:

$$p^*(c|y) = \left(1 / z^*(y,\eta) \right) \textstyle\prod_w p(y|c,w)^{\eta\, p(w)}\ . \ \square \tag{D.18}$$