Kikuchi approximations, Moebius decompositions, and variational principles

Alain Franc,
Michel Goulard,
& Nathalie Peyrard

Feb 26th, 2008

Contents

1 Introduction

In statistical physics, many distribution laws over a spatially explicit space can be written in an explicit form, but are untractable through direct calculation, as the number of operations required grows exponentially with the dimension of the system. This is the case for Boltzmann distribution

$$p\,(x) = \frac{1}{Z}\ \exp -\beta\, H(x) \tag{1.1}$$

over a lattice for the Ising model, where $H(x)$ is the energy of the system in state x and Z is the partition function $Z = \sum_x \exp -\beta\, H(x)$. $H(x)$ can in general be computed in polynomial time with the size of the system, whereas Z requires $\sum_x$ operations, which grows exponentially with n. Most of thermodynamic function can be derived from Z, hence computation of Z is crucial. One way to circumvent intractability is to approximate H by a function $\widetilde{H}$ as close as possible to H for which calculations are tractable. A natural example is $H(x) = \sum_i H_i(x_i)$, where $Z = \sum_x \left(\prod_i \exp -\beta\, H_i(x_i)\right) = \prod_i \left(\sum_{x_i} \exp -\beta\, H(x_i)\right)$. This amounts to approximating distribution $p(x)$ by a "simpler" (in a sense relevant for the expected calculations) distribution $q(x)$.

An untractable distribution p being given, there exist several ways to select the approximate distribution q. One way, not addressed specifically here, but worth being mentioned and compared with, is to select a family $\mathcal{F}$ of tractable distributions, compute the distribution $q \in \mathcal{F}$ as close as possible to p, and make calculation of desired quantities (partition function, marginal distribution etc.) on approximate distribution q. If the family is a family of univariate laws, $q(x) = \prod_i q_i(x_i)$, this yields mean field approximation. Pair and Bethe approximations are obtained by more complex calculations involving products of laws depending on singletons x_i and pairs (x_i, x_j). The way addressed here is Kikuchi approximation, or cluster variational methods (CVM). Kikuchi approximations at order r are expressed as products (at a given exponent) of marginal laws of p at order r at most. If $m_i(x_i)$ is the marginal law of p on i, Kikuchi approximation at order 1 simply is $\prod_i m_i(x_i)$. Note that in general, it is not the best approximation of p as a product $\prod_i q_i(x_i)$ (the latter being mean field approximation).

In this paper, we recall how Kikuchi approximations of a distribution p can

be retrieved as cut-off in so called Moebius decompositions of p at a given order, yielding an approximation q as a product $q = \prod_i q_i$ each function q_i being of lower order after cut-off. Intensively studied models such as graphical models are an example of these approximations (see [**?**]). A link between Kikuchi approximation and Moebius formula has been noticed by Schlijper (cited in [**?**]). Distributions q written as products $q = \prod_\alpha q_\alpha(x_\alpha)$ where $\alpha \subset \{1, n\}$ is a multi-index made of some nodes i only can be formalized by factor graphs: a factor graph is a bipartite graph of nodes i and factors α, with edge between i and α when i is in α (see [**?**]). Message passing on factor graphs is a technique to efficiently (i.e. in a reasonable time with reasonable memory) compute marginal probability distributions of q (see [**?** , chapter 26], [**?**]). It is exact when the topology of the graph is loop free (XXX preciser, erreur d'ecriture dans yediddia ..... qui parle du bipartite graph, qui est tjs loop free ...).

The name of Kikuchi is often associated to a technique which is related to, but different from, the previous question. Both types of references when used together (which is often meaningful) may create some confusion if proper references are not fully explicit. We try here to make them explicit. One way (among others) to quantify the distance between two distributions is to use the Kullback-Leibler distance

$$\mathrm{KL}(q\|p) = \sum_x q(x)\ \mathrm{Log}\ \frac{q(x)}{p\,(x)} \tag{1.2}$$

It is not a distance as it is not symmetric: $KL(q\|p\,) \neq KL(p\,\|q)$ in general. In the case where p is a Boltzmann distribution (or, a vocabulary of stochastic processes, a Gibbs distribution – therefore, it is necessary and sufficient that p is never equal to 0 –) i.e. $p\,(x) = (1/Z)\ \exp -\beta\, H(x)$, then $\mathrm{Log}\, p\,(x) = -\beta\, H(x) - \mathrm{Log}\, Z$ and it is classical to show that

$$\mathrm{KL}(q\|p) = \beta(\mathbb{F}(q) - \mathbb{F}(p)) \tag{1.3}$$

where the free energy $\mathbb{F}(p)$ of a distribution p is defined as $\beta\,\mathbb{F}(p) = -\mathrm{Log}\, Z$ or $\mathbb{F}(p) = \mathcal{U}(p) - \beta^{-1}\mathcal{H}(p)$. Gibbs inequality states that, p being given, $\mathrm{KL}(q\|p) \geq 0$ whatever the distribution q, with $\mathrm{KL}(q\|p) = 0$ for $q = p$. Then, $\mathbb{F}(q) \geq \mathbb{F}(p)$, and the distribution q the closest to p in the sense of the Kullback-Leibler distance is the one with minimum free energy. The quantity $\mathbb{F}(q)$ is called the Kikuchi free energy of the system. Then, approximating p by q is minimizing its Kikuchi free energy.

2 A simple calculation for a law with two variables

Let us have a set $\Omega = \Lambda \times \Lambda$, with $\Lambda = [1, q] \subset \mathbb{N}$, and p a probability law on Ω. The law p is completely determined by the values $p(x_1, x_2)$ with $x_1, x_2 \in \Lambda$. The marginal law m_1 on x_1 is given by

$$m_1(x_1) = \sum_{x_2} p(x_1, x_2) \tag{2.1}$$

and m_2 on x_2 by

$$m_2(x_2) = \sum_{x_1} p(x_1, x_2) \tag{2.2}$$

The question addressed here is to find a law q on Ω, the marginal of which are constrained to be equal respectively to give laws m_1 and m_2, and the entropy of which is maximum. It can be solved easily by Lagrange multipliers as in routine optimization under constraints. Here, we develop a direct calculation making use of convexity of function $x \operatorname{Log} x$. The reason for this choice is that the latter can easily be extended to general laws $p(x)$ with $x \in \Lambda^n$ under more general constraints on any marginal laws m_I with $I \subset \{1, n\}$, whereas the generalization of the calculation with Lagrange multipliers is less simple.

It is possible to write

$$p(x_1, x_2) = m_1(x_1)\, m_2(x_2)\, c(x_1, x_2) \tag{2.3}$$

which can be read as a definition of c. This is a general form, and any probability law $q(x_1, x_2)$ can be written this way. Independence between x_1 and x_2 yields $c = 1$. Then

$$\begin{aligned} m_1(x_1) &= \sum_{x_2} p(x_1, x_2) \\ &= \sum_{x_2} m_1(x_1)\, m_2(x_2)\, c(x_1, x_2) \\ &= m_1(x_1) \sum_{x_2} m_2(x_2)\, c(x_1, x_2) \end{aligned} \tag{2.4}$$

This yields

$$\sum_{x_2} m_2(x_2)\, c(x_1, x_2) = 1 \tag{2.5}$$

Similarly

$$\sum_{x_1} m_1(x_1)\, c(x_1, x_2) = 1 \tag{2.6}$$

Then

$$\begin{aligned}
\mathcal{H}(p) &= -\sum_{x_1,x_2} p\,(x_1, x_2)\ \mathrm{Log}\, p\,(x_1, x_2) \\
&= -\sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, c(x_1, x_2)\ \mathrm{Log}\,(m_1(x_1)\, m_2(x_2)\, c(x_1, x_2)) \\
&= \sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, c(x_1, x_2)\ (\mathrm{Log}\, m_1(x_1) + \mathrm{Log}\, m_2(x_2) + \mathrm{Log}\, c(x_1, x_2))
\end{aligned} \tag{2.7}$$

Now

$$\begin{aligned}
\sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, c(x_1, x_2)\ \mathrm{Log}\, m_1(x_1) &= \sum_{x_1} m_1(x_1)\ \mathrm{Log}\, m_1(x_1) \left(\sum_{x_2} m_2(x_2)\, c(x_1, x_2) \right) \\
&= \sum_{x_1} m_1(x_1)\ \mathrm{Log}\, m_1(x_1) \\
&= \mathcal{H}(m_1)
\end{aligned} \tag{2.8}$$

(as $\sum_{x_2} m_2(x_2)\, c(x_1, x_2) = 1$). Similarly

$$\sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, c(x_1, x_2)\ \mathrm{Log}\, m_2(x_2) = \mathcal{H}(m_2) \tag{2.9}$$

Then

$$\mathcal{H}(p) = \mathcal{H}(m_1) + \mathcal{H}(m_2) - \sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, c(x_1, x_2)\ \mathrm{Log}\, c(x_1, x_2) \tag{2.10}$$

Then, the entropy $\mathcal{H}(p)$ is maximum when the term due to interactions $\mathcal{I}(p) = \sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, c(x_1, x_2)\ \mathrm{Log}\, c(x_1, x_2)$ is minimum. It is now shown that $\forall\, p,\ \mathcal{I}(p) \geq 0$ and that the minimum $\mathcal{I} = 0$ is reached for $\forall\,(x_1, x_2),\ c(x_1, x_2) = 1$.

Let us notice that

$$\forall\, x \geq 0, \quad x\ \mathrm{Log}\, x \geq x - 1 \tag{2.11}$$

This is easily shown as $(x \text{ Log } x)' = 1 + \text{Log } x$, $(x \text{ Log } x)'' = 1/x > 0$, and $x \text{ Log } x$ is convex. It is always above $x-1$ as the slope at contact with $x-1$, at $x = 1$, is equal to 1. Then

$$\begin{aligned} \mathcal{I}(p) &\geq \sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, (c(x_1,x_2) - 1) \\ &= \sum_{x_1,x_2} m_1(x_1)\, m_2(x_2)\, c(x_1,x_2) - \sum_{x_1,x_2} m_1(x_1)\, m_2(x_2) - 1 \\ &= 1 - 1 \\ &= 0 \end{aligned} \tag{2.12}$$

As $\forall\ (x_1,x_2),\ m_1(x_1),\ m_2(x_2)\ c(x_1,x_2) > 0$, the value $\mathcal{I}(p) = 0$ is obtained at $\forall\ (x_1,x_2),\ c(x_1,x_2)\ \text{Log}\, c(x_1,x_2) = 0$. As $c(x_1,x_2) > 0$, this yields $\text{Log}\, c(x_1,x_2) = 0$ and $c(x_1,x_2) = 1$. Then, the maximum of $\mathcal{H}(p)$ is obtained at the minimum of $\mathcal{I}(p)$, which is obtained at $\mathcal{I}(p) = 0$ and

$$\mathcal{H}_{\max}(p) = \mathcal{H}(m_1) + \mathcal{H}(m_2) \tag{2.13}$$

This yields the classical result: among all the probability laws with constraints marginal laws (m_1, m_2), the one with maximum entropy is the one with independence $p = m_1 m_2$. This is a special (and easy) case of a more general result, which is developed here along he same way. Before that, we simply recall the classical extension to a probability law on a discrete space of size n.

3 Moebius inversion formula and decomposition

The Moebius inversion formula states that,

Let $\mathcal{V} = \{1, n\} \subset \mathbb{N}$. Let p be a probability law on $\Omega = \Lambda^n$. The set $\mathcal{P}(\mathcal{B})$ is in this example the partially ordered set for application of Moebius inversion formula.

Let $\mathcal{B} \subset \mathcal{V}$. Let us note $p|_{\mathcal{B}}$ the marginal law of p on $\mathcal{B}$. Then, there exist a set of maps $c_{\mathcal{A}}$ for any subset $\mathcal{A} \subset \mathcal{V}$

$$\begin{aligned} c_{\mathcal{A}} \ :\ \mathcal{A} &\longrightarrow \quad \mathbb{R}^+ \\ x|_{\mathcal{A}} &\longrightarrow c_{\mathcal{A}}(x|_{\mathcal{A}}) \end{aligned} \tag{3.1}$$

such that

$$\forall\, \mathcal{B} \subset \mathcal{V}, \quad p|_{\mathcal{B}}(x|_{\mathcal{B}}) = \prod_{\mathcal{A} \subset \mathcal{B}} c_{\mathcal{A}}(x|_{\mathcal{A}}) \tag{3.2}$$

In particular, if $\mathcal{B} = \mathcal{V}$

$$p\,(x) = \prod_{\mathcal{A} \subset \mathcal{V}} c_{\mathcal{A}}(x|_{\mathcal{A}}) \tag{3.3}$$

This is shown by constructing the maps $c_{\mathcal{A}}$. For sake of simplicity, if $\mathcal{B} = \{i\}$, we note $c_{\mathcal{B}}(x|_{\mathcal{B}}) \doteq c_i(x_i)$ and so forth for higher orders. Similarly, we note $m_i(x_i) \doteq p|_{\{i\}}$. Putting $\mathcal{B} = \{i\}$ yields

$$m_i(x_i) = c_i(x_i) \tag{3.4}$$

Then, all maps c_i are determined this way. Knowing that, and putting $\mathcal{B} = \{i, j\}$ yields

$$m_{ij}(x_i, x_j) = c_i(x_i)\, c_j(x_j)\, c_{ij}(x_i, x_j) \tag{3.5}$$

Thus

$$c_{ij}(x_i, x_j) = \frac{m_{ij}(x_i, x_j)}{m_i(x_i)\, m_j(x_j)} \tag{3.6}$$

Ordering these calculations by increasing order of I yields all the maps $c_{\mathcal{A}}$. Writing all the maps $c_{\mathcal{A}}$ as functions of marginal probabilities m_I as in equation 3.6 is a Moebius inversion formula. It yields that

$$c_I(x_I) = \prod_{I,J\,:\,|J|<|I|} m_J^{\mu_{I,J}}(x_J) \tag{3.7}$$

In 3.6, $I = (i, j)$, $\mu_{ij} = 1$, $\mu_i = \mu_j = -1$.

For this inversion to be possible, it is not necessary that all the subsets $\mathcal{B} \subset \mathcal{V}$ are considered. Let us have a subset $\mathcal{R}$ of all subsets $\mathcal{P}(\mathcal{V})$ of $\mathcal{V}$: $\mathcal{R} \subset \mathcal{P}(\mathcal{V})$. Under which conditions is it possible to derive a Moebius inversion formula on $\mathcal{R}$? Let us take as an example the set $\mathcal{V} = \{1, 2, 3, 4, 5\}$, and

$$\mathcal{R} = \{r_1 = \{1\},\ r_2 = \{2, 3\},\ r_3 = \{1, 4, 5\},\ r_4 = \{2, 3, 4, 5\}\} \tag{3.8}$$

We have

$$r_1 \subset r_3, \quad r_2 \subset r_4 \tag{3.9}$$

Then, let us assume that

$$\begin{aligned} p\left(x_1,x_2,x_3,x_4,x_5\right) = c_1(x_1)\, c_{23}(x_2,x_3)\, c_{145}(x_1,x_4,x_4) \;\times \ldots \\ \ldots \times c_{2345}(x_2,x_3,x_4,x_5)\, c_{12345}(x_1,x_2,x_3,x_4,x_5) \end{aligned} \tag{3.10}$$

Then

$$\begin{aligned} c_1(x_1) &= m_1(x_1) \\ c_{23}(x_2,x_3) &= m_{23}(x_2,x_3) \\ c_{145}(x_1,x_4,x_4) &= \frac{m_{145}(x_1,x_4,x_4)}{m_1(x_1)} \\ c_{2345}(x_2,x_3,x_4,x_5) &= \frac{m_{2345}(x_2,x_3,x_4,x_5)}{m_{23}(x_2,x_3)} \\ c_{12345}(x_1,x_2,x_3,x_4,x_5) &= \frac{p_{12345}(x_1,x_2,x_3,x_4,x_5)}{m_1(x_1)\, m_{23}(x_2,x_3)\, \frac{m_{145}(x_1,x_4,x_4)}{m_1(x_1)}\, \frac{m_{2345}(x_2,x_3,x_4,x_5)}{m_{23}(x_2,x_3)}} \\ &= \frac{p_{12345}(x_1,x_2,x_3,x_4,x_5)}{m_{145}(x_1,x_4,x_4)\; m_{2345}(x_2,x_3,x_4,x_5)} \end{aligned} \tag{3.11}$$

A Moebius inversion can be derived on any subset $\mathcal{R} \subset \mathcal{P}(\mathcal{V})$ of all the subsets of $\mathcal{V}$. No specific condition is required therefore.

Two sets $\mathcal{R}$ are often considered as they yield to classical decomposition. If $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a connected graph, the set $\mathcal{R}$ of connected subgraphs of $\mathcal{G}$ is closed by inclusion. Applying Moebius inversion formula on this subset yields the classical decompostion (XXX Pearl)

$$p\left(x\right) = \frac{\prod_{i\sim j} m_{ij}(x_i,x_j)}{\prod_k m_k^{d_k-1}(x_k)}\; c(x) \tag{3.12}$$

A second choice is the set of cliques in $\mathcal{G}$. A clique $\mathcal{C}$ in a graph $\mathcal{G}$ is a set of vertices such that there is an edge in $\mathcal{E}$ between any pair of vertices in the clique. The set of cliques is closed by inclusion too. This set is naturally connected to graphical models. In a graphical model, a potential $\phi_{\mathcal{C}}(x_{\mathcal{C}})$ is defined on each clique $\mathcal{C}$, and the global potential (as $\phi = \exp -\beta E$ where E is the energy in Ising model) is defined as

$$\phi(x) = \prod_{\mathcal{C}} \phi_{\mathcal{C}}(x_{\mathcal{C}}) \tag{3.13}$$

A key question is to compute a probability function

$$p\left(x\right)=\frac{1}{Z}\,\phi(x),\qquad Z=\sum_{x}\phi(x) \tag{3.14}$$

Computing Z is often untractable. Therefore, approximations are considered, and among the Kikuchi approximation.

4 Kikuchi approximation

A Kikuchi approximation $\mathcal{K}_r(p)$ of probability law p at order r is a cut-off of of development 3.3 at subsets $\mathcal{A}$ of a given size

$$\mathcal{K}_r(p)=\prod_{\mathcal{A}\,:\,|A|\leq r} c_{\mathcal{A}}(x|_{\mathcal{A}}) \tag{4.1}$$

For example, the Kikuchi approximation at order $r=2$ reads

$$\mathcal{K}_2(p)=\frac{\prod_{ij} m_{ij}(x_i,x_j)}{\prod_k m_k^{n-1}(x_k)} \tag{4.2}$$

This approximation may, or may not, be a probability law. Indeed, it may, or may not, sum up to one.

A Kikuchi approximation at order r is a function of marginal laws of p up to order r. It is in general, if I is a multi-index ($I=(i_1,\ldots,i_k)$, $k\in\{1,r\}$), with $|I|=k$, of the form

$$\mathcal{K}_r(p)=\prod_{|I|} m_I^{\alpha_I}(x|_I) \tag{4.3}$$

Let us call $A_p^r(x)$ the probability associated with $\mathcal{K}_r(p)$, that is $A_p^r(x)=\alpha\mathcal{K}_r(p)$ where α is a normalization constant. Then, it is possible to write

$$p\left(x\right)=A_p^r(x)\,C_p^r(x) \tag{4.4}$$

with $C_p^r(x)$ being made of maps $c_{\mathcal{A}}$ of order $s>r$.

5 Result

We show here that A_p^r is the probability law with maximal entropy among those the marginal laws of which are constraint to be the laws m_I . In other words, if $m_I(p)$ is the marginal law of p on multiindex I

$$m_I(q) = m_I(p) \quad \forall\, I \,:\, |I| \leq r \qquad \Rightarrow \qquad \mathcal{H}(q) \leq \mathcal{H}(A_p^r) \tag{5.1}$$

Let us suppose that q has same marginal laws than p up to order r. then, we can write

$$q(x) = A_p^r(x) C(x) \tag{5.2}$$

with $C(x)$ being a function of maps of order $\geq r$. It is then possible to develop

$$\begin{aligned}
\mathcal{H}(q) &= -\sum_x q\,(x)\ \mathrm{Log}\ q\,(x) \\
&= -\sum_x A_p^r(x)\, C(x)\ \mathrm{Log}\ A_p^r(x)\, C(x) \\
&= -\sum_x A_p^r(x)\, C(x)\ (\mathrm{Log}\ A_p^r(x) + \mathrm{Log}\, C(x)) \\
&= -\sum_x A_p^r(x)\, C(x)\ \mathrm{Log}\ A_p^r(x) - \sum_x A_p^r(x)\, C(x)\ \mathrm{Log}\, C(x)
\end{aligned} \tag{5.3}$$

As $C(x)\ \mathrm{Log}\, C(x) \geq C(x) - 1$, this yields

$$\mathcal{H}(q) \leq -\sum_x A_p^r(x)\, C(x)\ \mathrm{Log}\ A_p^r(x) - \sum_x A_p^r(x)(C(x) - 1) \tag{5.4}$$

As $A_p^r(x)$ is a probability law, $\sum_x A_p^r(x) = 1$, and as $q\,(x) = A_p^r(x)\, C(x)$, $\sum_x A_p^r(x)\, C(x) = 1$. Thus, $\sum_x A_p^r(x)(C(x) - 1) = 0$, and

$$\mathcal{H}(q) \leq -\sum_x A_p^r(x)\, C(x)\ \mathrm{Log}\ A_p^r(x) \tag{5.5}$$

We now show that

$$\begin{aligned}
\sum_x A_p^r(x)\, C(x)\ \mathrm{Log}\ A_p^r(x) &= \sum_x A_p^r(x)\ \mathrm{Log}\ A_p^r(x) \\
&= \mathcal{H}(A_p^r)
\end{aligned} \tag{5.6}$$

which shows that

$$\mathcal{H}(q) \leq \mathcal{H}(A_p^r) \tag{5.7}$$

We recall that

$$A_p^r(x) = \alpha * \prod_I m_I^{\alpha_I}(x) \tag{5.8}$$

where α is a normalizing constant, equal to 1 if $\mathcal{K}_r(p)$ is a probability (for sake of simplification we omit that in fact $m_I(x) = m_I(x|_I)$). Then

$$\text{Log}\, A_p^r(x) = \text{Log}\alpha + \sum_I \alpha_I \,\text{Log}\, m_I(x) \tag{5.9}$$

and

$$\sum_x A_p^r(x)\, C(x)\, \text{Log}\, A_p^r(x) = \text{Log}\alpha + \sum_I \alpha_I \sum_x A_p^r(x)\, C(x) \,\text{Log}\, m_I(x) \tag{5.10}$$

We now show that

$$\forall\, I, \quad \sum_x A_p^r(x)\, C(x)\, \text{Log}\, m_I(x) = \sum_x A_p^r(x)\, \text{Log}\, m_I(x) \tag{5.11}$$

Therefore, let us note J the multi-index of indices in $\mathcal{V} = \{1, n\}$ and not in I. Let us note y the subset $y = x|_I$ and $z = x|_J$. Then, $\sum_x = \sum_z \sum_y$. As $m_I(z)$ is the marginal law of q for I, we have, by definition

$$\begin{aligned} m_I(z) &= \sum_y q(z, y) \\ &= \sum_y A_p^r(z, y) C(z, y) \\ &= m_I(z) \sum_y \frac{A_p^r(z, y) C(z, y)}{m_I(z)} \end{aligned} \tag{5.12}$$

Then

$$\sum_y \frac{A_p^r(x) C(x)}{m_I(z)} = 1 \tag{5.13}$$

and

$$\begin{aligned}\sum_x A_p^r(z,y)\,C(z,y)\,\mathrm{Log}\,m_I(z) &= \sum_x \frac{A_p^r(z,y)\,C(z,y)}{m_I(z)}\; m_I(z)\,\mathrm{Log}\,m_I(z)\\ &= \sum_z m_I(z)\,\mathrm{Log}\,m_I(z)\sum_y \frac{A_p^r(z,y)\,C(z,y)}{m_I(z)}\\ &= \sum_z m_I(z)\,\mathrm{Log}\,m_I(z)\end{aligned} \tag{5.14}$$

Then

$$\begin{aligned}\sum_x A_p^r(x)\,C(x)\;\mathrm{Log}\,A_p^r(x) &= \mathrm{Log}\alpha + \sum_I \alpha_I \sum_x A_p^r(x)\,C(x)\,\mathrm{Log}\,m_I(x)\\ &= \mathrm{Log}\alpha + \sum_I \alpha_I \sum_z m_I(z)\,\mathrm{Log}\,m_I(z)\end{aligned} \tag{5.15}$$

In this form, it is possible to recognize $\mathcal{H}(A_p^r) = -\sum_x A_p^r(x)\,\mathrm{Log}\,(A_p^r(x))$. Indeed, the entropy $\mathcal{H}(p)$ of a law p is the expectation of $-\mathrm{Log}\,p$, as

$$\begin{aligned}\mathcal{H}(p) &= -\sum_x p\,(x)\,\mathrm{Log}\,p\,(x)\\ &= \mathbb{E}_p(-\mathrm{Log}\,p)\end{aligned} \tag{5.16}$$

where $\mathbb{E}_p(f(x)) = \sum_x p\,(x)\;f(x)$. Then

$$\begin{aligned}\mathcal{H}(A_p^r) &= \mathbb{E}_{A_p^r}(-\mathrm{Log}\,A_p^r)\\ &= -\mathrm{Log}\alpha - \mathbb{E}_{A_p^r}\left(\mathrm{Log}\prod_I m_I^{\alpha_I}\right)\\ &= -\mathrm{Log}\alpha - \mathbb{E}_{A_p^r}\left(-\sum_I \alpha_I\,\mathrm{Log}\,m_I\right)\\ &= -\mathrm{Log}\alpha - \sum_I \alpha_I \mathbb{E}_{A_p^r}(\mathrm{Log}\,m_I)\\ &= -\mathrm{Log}\alpha - \sum_I \alpha_I \sum_x m_I(x)\,\mathrm{Log}\,m_I(x)\end{aligned} \tag{5.17}$$

Indeed, recalling that $y \doteq x|_I$ and $z \doteq x|_J$

$$\begin{aligned} \mathbb{E}_{A_p^r}(-\mathrm{Log}\, m_I) &= -\sum_x A_p^r(x)\, \mathrm{Log}\, m_I(z) \\ &= -\sum_z \mathrm{Log}\, m_I(z) \sum_y A_p^r(z,y) \end{aligned} \tag{5.18}$$

and by definition $\sum_y A_p^r(z,y) = m_I(z)$.

6 Just for fun

$$\begin{aligned} \mathrm{KL}(q\|p) &= \sum_x q(x)\ \mathrm{Log}\ \frac{q(x)}{p\,(x)} \\ &= \mathbb{E}_q\ \mathrm{Log}\ q(x) - \mathbb{E}_q\ \mathrm{Log}\ p\,(x) \\ &= \mathbb{E}_q\ \mathrm{Log}\ q(x) + \mathbb{E}_q(\beta\, H(x) + \mathrm{Log}\ Z) \\ &= -\mathcal{H}(q) + \beta\, \mathcal{U}(q) + \mathrm{Log}\ Z \\ &= \beta\, \big(\mathcal{U}(q) - \beta^{-1}\mathcal{H}(q)\big) + \mathrm{Log}\ Z \\ &= \beta\, \mathbb{F}(q) + \mathrm{Log}\ Z \\ &= \beta\, (\mathbb{F}(q) - \mathbb{F}(p)) \end{aligned} \tag{6.1}$$