1 Correlated mutation rate

System of N indepedent spins $s_i \in \{-1, +1\}$, with fields h_i, such that at equilibrium

$$P_{eq}(\mathbf{s}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i s_i\right) = \prod_{i=1}^{N} \frac{e^{h_i s_i}}{Z_i}. \tag{1}$$

Starting from a sequence $\mathbf{s^0}$, the dynamic will be the following: at each time step, a unique position $1 \leq i \leq N$ is chosen at random with probability $1/N$, and spin s_i is re-emitted with equilibrium probability $e^{h_i s_i}/Z_i$. This dynamic will obviously lead to equilibrium, with some characteristic time proportional to N. We ask what the marginal distribution of spins $f_i(s_i)$ and $f_{ij}(s_i, s_j)$ are at time $t \ll N$.

Single site frequencies have the following distribution at time t

$$f_i^t(s_i|s_i^0) = \nu_1^t \delta(s_i = s_i^0) + (1 - \nu^t)\frac{e^{h_i s_i}}{Z_i}, \tag{2}$$

where $\nu_1 = 1 - 1/N = (N-1)/N$. To the order one in t/N, this has the following form

$$f_i^t(s_i|s_i^0) = \left(1 - \frac{t}{N}\right)\delta(s_i^0) + \frac{t}{N}\frac{e^{h_i s_i}}{Z_i}. \tag{3}$$

A naïve inference considering configurations distant of t from $\mathbf{s}^0$ as indepdant would give the following fields:

$$\begin{aligned} h_i^{inf}(t) &= \frac{1}{2}\log\left(\frac{f_i^t(1)}{1 - f_i^t(1)}\right) \\ &= -\frac{s_i^0}{2}\log\frac{t}{N}\frac{e^{-h_i s_i^0}}{Z_i}. \end{aligned} \tag{4}$$

Note: This seems to works numerically, up to $t \sim 2N$ on a $N = 32$ spins system with fields distributed normally and any starting configuration. In the sense that up to $t \sim 2N$, inferred fields are closer to equation (4) than to actual fields of the model.

For pairwise frequencies, one defines $\nu_2 = 1 - 2/N = (N-2)/N$. There are three possibilities for mutations at sites i and j, giving three terms in f_{ij}^t:

(i) Neither i nor j have been mutated. Corresponding probability is ν_2^t.

(ii) i has been mutated at least once, but not j. Corresponding probability is $\nu_1^t(1 - (1 - 1/(N-1))^t) = \nu_1^t - \nu_2^t$. Of course, the symetrical event is also possible, and situation *(ii)* should count two times.

(iii) Both i and j have been mutated at least once. By substraction of *(i)* and $2\cdot$*(ii)*, corresponding probability is $1 - 2\nu_1^t + \nu_2^t$.

Combining these three events, we have the following equation for pairwise frequencies:

$$f_{ij}^t(s_i, s_j|s_i^0, s_j^0) = \nu_2^t \delta(s_i^0, s_j^0) + (\nu_1^t - \nu_2^t)\left(\frac{e^{h_i s_i}}{Z_i}\delta(s_j^0) + \frac{e^{h_j s_j}}{Z_j}\delta(s_i^0)\right) + \left(1 - 2\nu_1^t + \nu_2^t\right)\left(\frac{e^{h_i s_i}e^{h_j s_j}}{Z_i Z_j}\right). \tag{5}$$

If one does not know the starting sequence s^0, it is necessary to integrate over all its possible values, with probability distribution $P_{eq}(s^0)$. It can also be interesting to take into account finite size sampling effects. If frequency f_i is obtained through independently repeating the dynamic described above M times, always starting from s^0, then one can expect a finite sample error effect having a variance

$$\sigma_{f1}^2(s_i|s^0) = f_i(s_i|s^0)\left(1 - f_i(s_i|s^0)\right)$$

In this scenario, the frequency measured for a given tree with root s^0 and M branches of length t with configurations $\sigma^1 \ldots \sigma^M$, is the following random variable:

$$\tilde{f}_i(s_i|s^0)[s^0, u] = \frac{1}{M}\sum_{a=1}^{M}\delta(\sigma_i^a = s_i) \equiv \nu_1^t \delta(s_i = s_i^0) + (1 - \nu^t)P_{eq}(s_i) + u(s_i|s^0) \tag{6}$$

with gaussian random variable $u \sim \mathcal{N}\left(0, \sigma_{f1}^2(s_i|s^0)/M\right)$.

2 Uncorrelated mutations

Similar process as in the previous section, but mutations can now take place at two sites at the same time in an uncorrelated manner. At each time step, each site has a probability μ to mutate, being re-emitted with equilibrium probability.
If the tree of configurations is star like, with all configurations originating from a single root $\mathbf{s}^0$ with no intermediate branching, no correlations can be generated by this process. However, it is possible to give rise to correlations if configurations originate from two different roots $\mathbf{s}^1$ and $\mathbf{s}^2$, as would happen in a binary tree for instance. If one considers children configurations of $\mathbf{s}^1$ and $\mathbf{s}^2$ as a single sample, the single and two sites probability frequencies can be written in following way:

$$P(s_i|\mathbf{s}^1,\mathbf{s}^2) = \frac{1}{2}\sum_{a=1}^{2}\left[\nu^t\delta(s_i=s_i^1)+(1-\nu^t)P_{eq}(s_i)\right], \tag{7}$$

and

$$\begin{aligned} P(s_i,s_j|\mathbf{s}^1,\mathbf{s}^2) = \frac{1}{2}\sum_{a=1}^{2}\big[&\nu^{2t}\delta(s_i=s_i^a,s_j=s_j^a)\\ &+\nu^t(1-\nu^t)\left(\delta(s_i=s_i^a)P_{eq}(s_j)+\delta(s_j=s_j^a)P_{eq}(s_i)\right)\\ &+(1-\nu^t)^2P_{eq}(s_i,s_j)\big], \end{aligned} \tag{8}$$

where $\nu = 1-\mu$.
Corresponding connected correlations can be computed exactly from these two equations. With shorter notations $\delta_i(\sigma) = \delta(s_i=\sigma)$, they read

$$c_{ij}(s_i,s_j|\mathbf{s}^1,\mathbf{s}^2) = \frac{1}{4}\left[\delta_i(s_i^1)\delta_j(s_j^1)+\delta_i(s_i^2)\delta_j(s_j^2)-\delta_i(s_i^1)\delta_j(s_j^2)-\delta_i(s_i^2)\delta_j(s_j^1)\right]\nu^{2t}. \tag{9}$$

If configurations $\mathbf{s}^1$ and $\mathbf{s}^2$ both originate from the same configuration $\mathbf{s}^0$, it is reasonnable to consider that they are independent from each other – in the sense that $P(s_i^1,s_j^2) = P(s_i^1)P(s_j^2)$ – and identically distributed. However, it is possible that they are not distributed according to P_{eq}, for instance if the time separating them from $\mathbf{s}^0$ is too short, or if their distribution is known from a bigger tree of which $\mathbf{s}^0$ is not the root, thus introducing correlations. From these assumptions, one can compute the average value of the correlation over the root sequences:

$$\langle c_{ij}(s_i,s_j|\mathbf{s^0})\rangle = \frac{1}{2}\left(P^1(s_i,s_j|\mathbf{s}^0)-P^1(s_i|\mathbf{s}^0)P^1(s_j|\mathbf{s}^0)\right)\nu^{2t} \tag{10}$$

where P^1 represents the distribution of configurations $\mathbf{s}^1$ and $\mathbf{s}^2$, (*i.e.* $P^1(s_i) = P(s_i^a = s_i)$, $a = 1$ ou 2). The average value of correlation at the second step of the tree, that is for configurations below $\mathbf{s}^1$ and $\mathbf{s}^2$, is the value of correlation at the superior step with $\mathbf{s}^0$ fixed. If parent configurations are distributed with equilibrium probability P_{eq}, the average correlation is zero.
One can also compute the standard deviation of the correlations, which gives an indication of how large the expected correlation would be on average, for a given distribution of root configurations. If $\mathbf{s}^1$ and $\mathbf{s}^2$ are distributed according to P_{eq}, one can show that

$$\langle c_{ij}(s_i,s_j)^2\rangle = \frac{\nu^{4t}}{4}P_{eq}(s_i)P_{eq}(s_j)(1-P_{eq}(s_i))(1-P_{eq}(s_j)). \tag{11}$$