1 Autoregressive Neural Networks

Consider, as the simplest non-trivial[1] example a three-qubit system acted upon by a single-qubit gate on the third qubit. New probabilities are give by

$$\begin{aligned} P'(a_1, a_2, a_3) &= \mathcal{O}^{a_3 a_2 a_1}_{b_3 b_2 b_1} \delta_{(a_2, b_2)} \delta_{(a_1, b_1)} P(b_1, b_2, b_3) \\ &= \mathcal{O}^{a_3}_{b_3} P(a_1, a_2, b_3) \end{aligned}$$

The autoregressive property ensures that the conditional probabilities, in which the joint probability can be decomposed, are implemented by seperate neural networks. In probabilities, the following conditions hold

$$\begin{aligned} P(a_1, a_2, a_3) &= P(a_1) P(a_2|a_1) P(a_3|a_2 a_1), \\ P'(a_1, a_2, a_3) &= P(a_1) P(a_2|a_1) P'(a_3|a_2 a_1). \end{aligned}$$

Hence we can write

$$P'(a_3|a_2 a_1) = \mathcal{O}^{a_3}_{b_3} P(b_3|a_2 a_1) \tag{1}$$

Assuming the conditional probabilites for the measurement outcome on qubit three are encooded by 1-layer feed-forward networks with softmax activation function allows us to rewrite 1 as

$$\frac{\exp\left(\beta'^3_{a_3} + \gamma'^{(23)}_{a_3 a_2} + \gamma'^{(13)}_{a_3 a_1}\right)}{\sum_{i,j,k} \exp\left(\beta'^3_i \gamma'^{(23)}_{ij} + \gamma'^{(13)}_{i,}\right)} = \sum_{b_3} \mathcal{O}^{a_3}_{b_3} \frac{\exp\left(\beta'^3_{b_3} + \gamma'^{(23)}_{b_3 a_2} + \gamma'^{(13)}_{b_3 a_1}\right)}{\sum_{i,j,k} \exp\left(\beta'^3_i \gamma'^{(23)}_{ij} + \gamma'^{(13)}_{i,}\right)}$$

Ignoring the LHS normlization, we can solve for $\{\beta'^3, \gamma'^{(23)}, \gamma'^{(13)}\}$. This gives

$$\beta'^3 + \gamma'^{(23)} + \gamma'^{(13)} = \ln\left(\sum_{b_3} \mathcal{O}^{a_3}_{b_3} \frac{\exp\left(\beta'^3_{b_3} + \gamma'^{(23)}_{b_3 a_2} + \gamma'^{(13)}_{b_3 a_1}\right)}{\sum_{i,j,k} \exp\left(\beta'^3_i \gamma'^{(23)}_{ij} + \gamma'^{(13)}_{i,}\right)}\right) \tag{2}$$

We now see that a fundamental formal problem arises: The new biases and weight matrices $\{\beta'^3, \gamma'^{(23)}, \gamma'^{(13)}\}$ all can be though of as functions mapping an input (a_1, a_2, a_3) to some number. The fact that the parameters are summed however indicates a function of the structure

$$\phi = f(a_3) + g(a_3, a_2) + h(a_3, a_1). \tag{3}$$

Because of this structure the function can only depend on 32 distinct values (no $a_1 a_2$-coupling). But the right hand side in general is of the form

$$\phi' = f'(a_1, a_2, a_3) \tag{4}$$

This discrete function depends on *3-tuples*, whereas ϕ only depends on pairs of *2-tuples*. In general, ϕ' can take more distinct values than ϕ, hence justifying that there cannot exist some $\{\beta'^3, \gamma'^{(23)}, \gamma'^{(13)}\}$ such that $\phi = \phi'$.

[1] not overparametrized

How does this agree with the fact that for 2-qubit systems correct updates can be found? This is because for 2-qubit systems we just have 16 different (a_1, a_2)-pairs which can be explicitly considered since they all correspond to entries in the weight matrix $\gamma'^{(12)}$. For bigger systems, we can use the same approach to encode probabilities if we replace weight matrices by *weight tensors*. However, since these scale exponentially with system size (all possible outcomes need to be considered), this approach is not useful for the efficient simulation of quantum algorithms with neural networks.