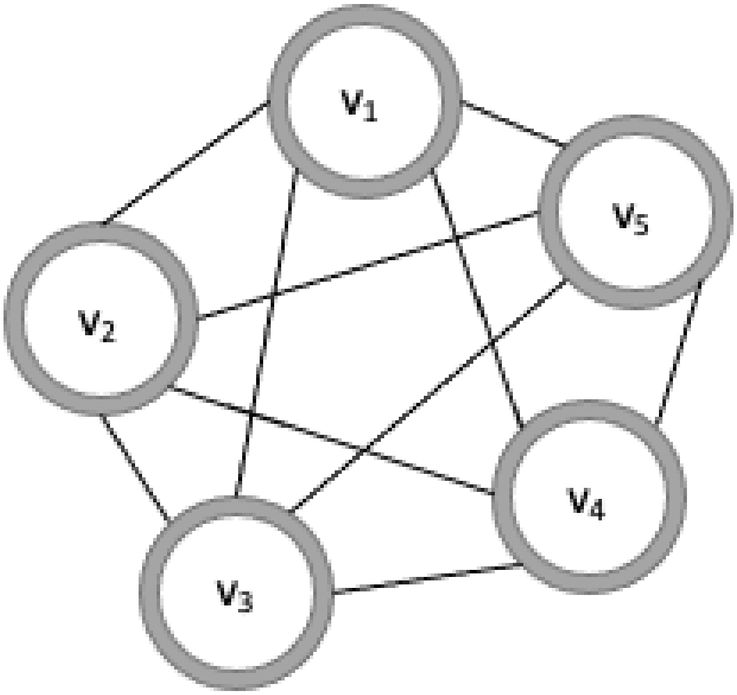
discrete_boltzmann

February 15, 2022

1 Discrete Boltzmann Machines

1.0.1 Discrete exponential distributions

Consider a D-dimensional space $\mathcal{V}$ of values. We notionally consider that each point $\mathbf{v} = (v_1, v_2, \ldots, v_D) \in \mathcal{V}$ is an instance of stochastic variables $(V_1, V_2, \ldots, V_D)$. A Boltzmann machine (BM) now takes the form of a completely connected, undirected graph, with a distinct node for each variable, as shown in the figure below.



For convenience, we now restrict ourselves to the case where $\mathcal{V}$ is a space of discrete values. Note that for continuous variables, (some of) the corresponding summations below would be replaced by integrations, although the resulting derivations will be of similar form.

A discrete Boltzmann machine now has an energy function $E : \mathcal{V} \rightarrow \mathbb{R}$ that induces the probability distribution

$$p(\mathbf{v}) \quad = \quad \frac{e^{-E(\mathbf{v})}}{\sum_{\mathbf{v}' \in \mathcal{V}} e^{-E(\mathbf{v}')}} \,. \tag{1}$$

In practice, it is more convenient to let $E = -f$ for some complementary function $f : \mathcal{V} \to \mathbb{R}$, which forms the exponential family

$$p(\mathbf{v}) \quad = \quad \frac{e^{f(\mathbf{v})}}{\sum_{\mathbf{v}' \in \mathcal{V}} e^{f(\mathbf{v}')}} \,. \tag{2}$$

1.0.2 Parameter estimation

We now suppose that $f(\mathbf{v})$ is implicitly parameterised by some collection of parameters, denoted by Θ. Our task is therefore to estimate Θ from a data-set of known training points. An obvious choice is to jointly maximise the likelihood $p(\mathbf{v})$ of each training point $\mathbf{v}$, under the assumption that training cases are independent. This is equivalent to maximising the log-likelihood, given by

$$\ln p(\mathbf{v}) \quad = \quad f(\mathbf{v}) - \ln \sum_{\mathbf{v}' \in \mathcal{V}} e^{f(\mathbf{v}')} \,. \tag{3}$$

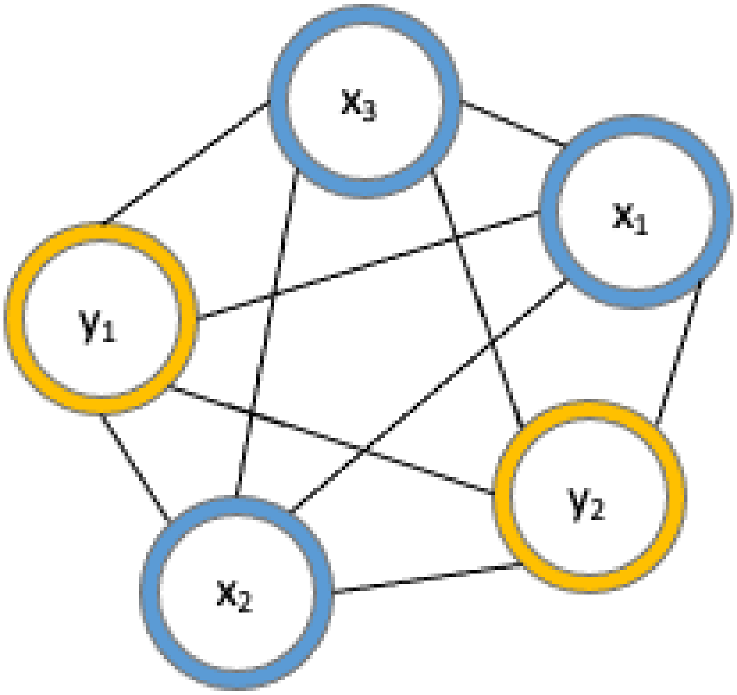
The usual choice for maximisation is gradient ascent, in one of its many variants. Hence, for some arbitrary parameter θ, the gradient of the log-likelihood is

$$\begin{aligned}
\nabla \ln p(\mathbf{v}) \quad &= \quad \nabla f(\mathbf{v}) - \nabla \ln \sum_{\mathbf{v}' \in \mathcal{V}} e^{f(\mathbf{v})} && (4) \\
&= \quad \nabla f(\mathbf{v}) - \frac{\sum_{\mathbf{v}' \in \mathcal{V}} e^{f(\mathbf{v}')} \, \nabla f(\mathbf{v}')}{\sum_{\mathbf{v}' \in \mathcal{V}} e^{f(\mathbf{v}')}} && (5) \\
&= \quad \nabla f(\mathbf{v}) - \sum_{\mathbf{v}' \in \mathcal{V}} p(\mathbf{v}') \, \nabla f(\mathbf{v}') && (6) \\
&= \quad \nabla f(\mathbf{v}) - \mathbb{E}_{\mathcal{V}} \left[\nabla f(\mathbf{v}') \right] \,. && (7)
\end{aligned}$$

The biggest problem in practice with Boltzmann machines is that $p(\mathbf{v})$ is intractable to compute in general, due largely to the *curse of dimensionality*. Hence, the unconditional expectation $\mathbb{E}_{\mathcal{V}}[\cdot]$ is also intractable. In the following sections, we discuss approximation techniques for handling this intractability.

1.0.3 Partitioned Boltzmann machine

The usual rationale for constructing a Boltzmann machine, or indeed for assuming any probability distribution, is for the purpose of prediction. Let us therefore suppose that the point $\mathbf{v} \in \mathcal{V}$ may be partitioned into two sub-points, $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. We also suppose that $\mathbf{x}$ and $\mathbf{y}$ may be *stitched* back together to obtain $\mathbf{v} = \check{\mathbf{v}}(\mathbf{x}, \mathbf{y})$. This *partitioned* Boltzmann machine is shown in the figure below.



For convenience, let us define $\breve{f}(\mathbf{x},\mathbf{y}) \doteq f(\breve{\mathbf{v}}(\mathbf{x},\mathbf{y})) = f(\mathbf{v})$. Then the joint distribution becomes

$$p(\mathbf{x},\mathbf{y}) \quad = \quad \frac{e^{\breve{f}(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')}}\,. \tag{8}$$

In addition, due to the non-directionality of edges, we may also predict in either direction, namely

$$p(\mathbf{y}\mid\mathbf{x}) \quad = \quad \frac{e^{\breve{f}(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x},\mathbf{y}')}}\,, \tag{9}$$

or

$$p(\mathbf{x}\mid\mathbf{y}) \quad = \quad \frac{e^{\breve{f}(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{x}'\in\mathcal{X}} e^{\breve{f}(\mathbf{x}',\mathbf{y})}}\,. \tag{10}$$

The corresponding marginal distributions are

$$p(\mathbf{y}) \quad = \quad \frac{\sum_{\mathbf{x}'\in\mathcal{X}} e^{\breve{f}(\mathbf{x}',\mathbf{y})}}{\sum_{\mathbf{y}'\in\mathcal{Y}}\sum_{\mathbf{x}'\in\mathcal{X}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')}}\,, \tag{11}$$

and

$$p(\mathbf{x}) \quad = \quad \frac{\sum_{\mathbf{y}\in\mathcal{Y}} e^{\breve{f}(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')}}\,, \tag{12}$$

respectively.

For convenience, we from now on suppose that the partitioned BM is treated as a predictive model with *input* $\mathbf{x}$ and *output* $\mathbf{y}$, although it will always be able to operate in reverse.

Before we proceed to the examination of parameter estimation for these various models, we first digress to the additional technique that we will utilise for handling the intractability of BMs, dicussed in the next section.

1.0.4 Mean field approximation

The mean field approximation states that the mean value of a function averaged over a number of points is approximately equal to the function evaluated at the average of those points. To demonstrate this, first let $\bar{\mathbf{v}} = \mathbb{E}_{\mathcal{V}}[\mathbf{v}]$ be the average of all the points in $\mathcal{V}$. Next, consider the first-order Taylor series approximation of some $g(\mathbf{v})$ about $\bar{\mathbf{v}}$, namely

$$g(\mathbf{v}) \quad \approx \quad g(\bar{\mathbf{v}}) + (\mathbf{v} - \bar{\mathbf{v}})^T \, \nabla g(\bar{\mathbf{v}}) \,. \tag{13}$$

Thus, taking the expectation over $\mathcal{V}$, it follows that

$$\begin{aligned} \mathbb{E}_{\mathcal{V}}[g(\mathbf{v})] \quad &\approx \quad g(\bar{\mathbf{v}}) + (\mathbb{E}_{\mathcal{V}}[\mathbf{v}] - \bar{\mathbf{v}})^T \, \nabla g(\bar{\mathbf{v}}) && (14) \\ &= \quad g(\bar{\mathbf{v}}) = g(\mathbb{E}_{\mathcal{V}}[\mathbf{v}]) \,. && (15) \end{aligned}$$

This is the mean field approximation (MFA).

If we proceed further to the second term in the Taylor series expansion (not shown here), then it becomes apparent that the accuracy of the approximation depends on both the smoothness of the function (especially its second derivative) and the variance of the points in $\mathcal{V}$. However, my experience is that MFA works very well in practice, especially for computing BM gradients that are otherwise intractable.

1.0.5 Joint likelihood optimisation

We suppose that the training data-set specifies both $\mathbf{x}$ and $\mathbf{y}$. Thus, we utilise the joint model $p(\mathbf{x}, \mathbf{y})$, defined in an earlier section. Observe that

$$\begin{aligned} \ln p(\mathbf{x},\mathbf{y}) \quad &= \quad \breve{f}(\mathbf{x},\mathbf{y}) - \ln \sum_{\mathbf{x}'\in\mathcal{X}} \sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')} && (16) \\ \Rightarrow \nabla \ln p(\mathbf{x},\mathbf{y}) \quad &= \quad \nabla\breve{f}(\mathbf{x},\mathbf{y}) - \frac{\sum_{\mathbf{x}'\in\mathcal{X}} \sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')} \, \nabla\breve{f}(\mathbf{x}',\mathbf{y}')}{\sum_{\mathbf{x}'\in\mathcal{X}} \sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')}} && (17) \\ &= \quad \nabla\breve{f}(\mathbf{x},\mathbf{y}) - \sum_{\mathbf{x}'\in\mathcal{X}} \sum_{\mathbf{y}'\in\mathcal{Y}} p(\mathbf{x}',\mathbf{y}') \, \nabla\breve{f}(\mathbf{x}',\mathbf{y}') && (18) \\ &= \quad \nabla\breve{f}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathcal{X},\mathcal{Y}}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right] \,. && (19) \end{aligned}$$

For convenience, we now let $\mathcal{X}' \equiv \mathcal{X}$ (and similarly for $\mathcal{Y}'$), where the prime distinguishes expectation over $\mathbf{x}' \in \mathcal{X}'$ from expectation over $\mathbf{x} \in \mathcal{X}$. Thus, we write

$$\begin{aligned} \nabla \ln p(\mathbf{x},\mathbf{y}) \quad &= \quad \nabla\breve{f}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathcal{X}',\mathcal{Y}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right] && (20) \\ &= \quad \nabla\breve{f}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathcal{X}'}\left[\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right] \,, && (21) \end{aligned}$$

where we have made use of the BM partitioning.

Note that by $\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}$ I really mean $\mathbb{E}_{\mathcal{Y}'|\mathbf{x}'}$, and thus $\mathbb{E}_{\mathcal{X}'}[\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}]$ really means $\mathbb{E}_{\mathbf{x}'\in\mathcal{X}'}[\mathbb{E}_{\mathcal{Y}'|\mathbf{x}'}]$. However, I didn't feel like mixing sets and points. Alternatively, I could have just written $\mathbb{E}_{\mathbf{y}'|\mathbf{x}'}$, as I have done in other notebooks, although this loses explicit mention of the domain. In my experience, all expectation notation suffers from exposing some explicit dependencies whilst hiding other implicit dependencies, and is thus never entirely unambiguous.

Now, computing $p(\mathbf{x}')$ is still intractable in general. However, we do know $\mathbf{x}$ and $\mathbf{y}$. Hence, we make further use of the partitioning by taking the approximation

$$\nabla\ln p(\mathbf{x},\mathbf{y}) \;\approx\; \nabla\breve{f}(\mathbf{x},\mathbf{y})-\mathbb{E}_{\mathcal{X}'|\mathcal{Y}}\left[\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right]\,. \tag{22}$$

In other words, we keep alternating between use of the predictive models $p(\mathbf{x}\mid\mathbf{y})$ and $p(\mathbf{y}\mid\mathbf{x})$ until we reach known values of the conditional. We shall call this the *conditional expectation approximation* (CEA), although another appropriate name would be *conditional expectation alternation* - take your pick.

Note that we assume that these predictive models are tractable to compute!

Next, we define the expectation functions

$$\bar{\mathbf{x}}(\mathbf{y}) \doteq \mathbb{E}_{\mathcal{X}|\mathcal{Y}}[\mathbf{x}]\,,\qquad \bar{\mathbf{y}}(\mathbf{x}) \doteq \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\mathbf{y}]\,. \tag{23}$$

These convenience functions allow us to more easily apply MFA, giving

$$\begin{aligned}
\nabla\ln p(\mathbf{x},\mathbf{y}) &\approx \nabla\breve{f}(\mathbf{x},\mathbf{y})-\mathbb{E}_{\mathcal{X}'|\mathcal{Y}}\left[\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right] &(24)\\
&\approx \nabla\breve{f}(\mathbf{x},\mathbf{y})-\mathbb{E}_{\mathcal{X}'|\mathcal{Y}}\left[\nabla\breve{f}(\mathbf{x}',\bar{\mathbf{y}}(\mathbf{x}'))\right] &(25)\\
&\approx \nabla\breve{f}(\mathbf{x},\mathbf{y})-\nabla\breve{f}(\bar{\mathbf{x}}(\mathbf{y}),\bar{\mathbf{y}}(\bar{\mathbf{x}}(\mathbf{y})))\,. &(26)
\end{aligned}$$

Alternatively, we may rewrite the gradient as

$$\begin{aligned}
\nabla\ln p(\mathbf{x},\mathbf{y}) &= \nabla\breve{f}(\mathbf{x},\mathbf{y})-\mathbb{E}_{\mathcal{Y}'}\left[\mathbb{E}_{\mathcal{X}'|\mathcal{Y}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right] &(27)\\
&\approx \nabla\breve{f}(\mathbf{x},\mathbf{y})-\mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\mathbb{E}_{\mathcal{X}'|\mathcal{Y}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right] &(28)\\
&\approx \nabla\breve{f}(\mathbf{x},\mathbf{y})-\mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\nabla\breve{f}(\bar{\mathbf{x}}(\mathbf{y}'),\mathbf{y}')\right] &(29)\\
&\approx \nabla\breve{f}(\mathbf{x},\mathbf{y})-\nabla\breve{f}(\bar{\mathbf{x}}(\bar{\mathbf{y}}(\mathbf{x})),\bar{\mathbf{y}}(\mathbf{x}))\,. &(30)
\end{aligned}$$

It is not clear which alternative is preferable. However, if we notionally think of $\mathbf{x}$ as the input, then this perhaps suggests the latter approxmation might be favoured. I have not implemented the former variant, but the latter variant appears to work well in practice.

In either case, we still cannot tractably compute the joint training score, $\ln p(\mathbf{x},\mathbf{y})$. However, one possible approach is to observe that

$$\begin{aligned}
p(\mathbf{x},\mathbf{y}) &= p(\mathbf{y}\mid\mathbf{x})\,p(\mathbf{x}) &(31)\\
&= p(\mathbf{y}\mid\mathbf{x})\sum_{\mathbf{y}'\in\mathcal{Y}'}p(\mathbf{y}')\,p(\mathbf{x}\mid\mathbf{y}') &(32)\\
&= p(\mathbf{y}\mid\mathbf{x})\,\mathbb{E}_{\mathcal{Y}'}\left[p(\mathbf{x}\mid\mathbf{y}')\right] &(33)\\
\Rightarrow p(\mathbf{x},\mathbf{y}) &\approx p(\mathbf{y}\mid\mathbf{x})\,\mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[p(\mathbf{x}\mid\mathbf{y}')\right] &(34)\\
&\approx p(\mathbf{y}\mid\mathbf{x})\,p(\mathbf{x}\mid\bar{\mathbf{y}}(\mathbf{x}))\,, &(35)
\end{aligned}$$

via CEA and MFA. Once again, this has been implemented and tested, and works well in practice.

1.0.6 Marginal likelihood optimisation

We now suppose that the training data-set only specifies $\mathbf{x}$ but not $\mathbf{y}$. Thus, we might utilise the marginal likelihood $p(\mathbf{x})$, defined in an earlier section. Observe that

$$\begin{aligned}
\ln p(\mathbf{x}) &= \ln \sum_{\mathbf{y}\in\mathcal{Y}} e^{\breve{f}(\mathbf{x},\mathbf{y})} - \ln \sum_{\mathbf{x}'\in\mathcal{X}} \sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')} &(36)\\
\Rightarrow \nabla \ln p(\mathbf{x}) &= \frac{\sum_{\mathbf{y}\in\mathcal{Y}} e^{\breve{f}(\mathbf{x},\mathbf{y})}\,\nabla\breve{f}(\mathbf{x},\mathbf{y})}{\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x},\mathbf{y}')}} - \frac{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')}\,\nabla\breve{f}(\mathbf{x}',\mathbf{y}')}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}')}} &(37)\\
&= \sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{y}\mid\mathbf{x})\,\nabla\breve{f}(\mathbf{x},\mathbf{y}) - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} p(\mathbf{x}',\mathbf{y}')\,\nabla\breve{f}(\mathbf{x}',\mathbf{y}') &(38)\\
&= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{\mathcal{X}',\mathcal{Y}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\,. &(39)
\end{aligned}$$

We observe that a general pattern has emerged here, namely that summation over a collection of variables in the log-likelihood corresponds, in the gradient, to the expectation over those same variables conditional on the remaining variables. Thus, $\sum_{\mathcal{Y}}$ in the first term on the right-hand side becomes $\mathbb{E}_{\mathcal{Y}|\mathcal{X}}$, and $\sum_{\mathcal{X}',\mathcal{Y}'}$ in the second term becomes $\mathbb{E}_{\mathcal{X}',\mathcal{Y}'}$.

Now, applying CEA gives

$$\nabla \ln p(\mathbf{x}) \;\approx\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\mathbb{E}_{\mathcal{X}'|\mathcal{Y}}\left[\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right]\right]\,. \tag{40}$$

Finally, applying MFA gives

$$\nabla \ln p(\mathbf{x}) \;\approx\; \nabla\breve{f}(\mathbf{x},\bar{\mathbf{y}}) - \nabla\breve{f}(\bar{\mathbf{x}}',\bar{\mathbf{y}}')\,, \tag{41}$$

where $\bar{\mathbf{y}} = \bar{\mathbf{y}}(\mathbf{x})$, $\bar{\mathbf{x}}' = \bar{\mathbf{x}}(\bar{\mathbf{y}}) = \bar{\mathbf{x}}(\bar{\mathbf{y}}(\mathbf{x}))$, and $\bar{\mathbf{y}}' = \bar{\mathbf{y}}(\bar{\mathbf{x}}') = \bar{\mathbf{y}}(\bar{\mathbf{x}}(\bar{\mathbf{y}}(\mathbf{x})))$. Note that the ordering of these (shorthand) computations corresponds to computing the expectations from left (outside) to right (inside). This is another general pattern.

This version of the gradient has been tested with a Bernoulli Restricted BM, and works well. The alternative version, namely

$$\begin{aligned}
\nabla \ln p(\mathbf{x}) &\approx \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\mathbb{E}_{\mathcal{X}'|\mathcal{Y}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right] &(42)\\
&\approx \nabla\breve{f}(\mathbf{x},\bar{\mathbf{y}}(\mathbf{x})) - \nabla\breve{f}(\bar{\mathbf{x}}(\bar{\mathbf{y}}(\mathbf{x})),\bar{\mathbf{y}}(\mathbf{x}))\,, &(43)
\end{aligned}$$

has not been tested. However, the presence of the same term $\bar{\mathbf{y}}(\mathbf{x})$ on both sides of the difference suggests that the reconstruction of $\mathbf{y}'$ would be poor!

Lastly, we observe that we cannot tractably compute the marginal score, $\ln p(\mathbf{x})$, of training case $\mathbf{x}$, for the same reason that we cannot in general compute the unconditional probability $p(\mathbf{x})$. However, we recall from above that

$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{x},\mathbf{y}) &(44)\\
&= \sum_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{x}\mid\mathbf{y})\,p(\mathbf{y}) &(45)\\
&= \mathbb{E}_{\mathcal{Y}}\left[p(\mathbf{x}\mid\mathbf{y})\right] &(46)\\
&\approx \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[p(\mathbf{x}\mid\mathbf{y})\right] &(47)\\
&\approx p(\mathbf{x}\mid\bar{\mathbf{y}}(\mathbf{x}))\,, &(48)
\end{aligned}$$

via CEA then MFA.

1.0.7 Conditional likelihood optimisation

Lastly, we look at the case where we wish to directly optimise the predictive model $p(\mathbf{y} \mid \mathbf{x})$ instead of the joint likelihood $p(\mathbf{x}, \mathbf{y})$. Assuming we know both $\mathbf{x}$ and $\mathbf{y}$, then (from our earlier derivation) we have

$$\begin{aligned}
\ln p(\mathbf{y} \mid \mathbf{x}) &= \breve{f}(\mathbf{x}, \mathbf{y}) - \ln \sum_{\mathbf{y}' \in \mathcal{Y}} e^{\breve{f}(\mathbf{x}, \mathbf{y}')} && (49) \\
\Rightarrow \nabla \ln p(\mathbf{y} \mid \mathbf{x}) &= \nabla \breve{f}(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{\mathcal{Y}' \mid \mathcal{X}} \left[\nabla \breve{f}(\mathbf{x}, \mathbf{y}') \right] && (50) \\
&\approx \nabla \breve{f}(\mathbf{x}, \mathbf{y}) - \nabla \breve{f}(\mathbf{x}, \bar{\mathbf{y}}(\mathbf{x})) \,, && (51)
\end{aligned}$$

via MFA. The log-likelihood score, $\ln p(\mathbf{y} \mid \mathbf{x})$, can now be computed directly from the predictive model.

I should add, as a note of caution obtained from testing various other discriminative models, that directly optimising the conditional predictive model can often exacerbate the effect of over-training, although this depends strongly on the training data.

Also note that there is no explicit modelling of the distribution of $\mathbf{x}$, and hence any related model parameters required for computing $p(\mathbf{x} \mid \mathbf{y})$ would need to be estimated in some other fashion. We could, for example, alternate between the gradient updates $\nabla \ln p(\mathbf{y} \mid \mathbf{x})$ and $\nabla \ln p(\mathbf{x} \mid \mathbf{y})$, where

$$\begin{aligned}
\ln p(\mathbf{x} \mid \mathbf{y}) &= \breve{f}(\mathbf{x}, \mathbf{y}) - \ln \sum_{\mathbf{x}' \in \mathcal{X}} e^{\breve{f}(\mathbf{x}', \mathbf{y})} && (52) \\
\Rightarrow \nabla \ln p(\mathbf{x} \mid \mathbf{y}) &= \nabla \breve{f}(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{\mathcal{X}' \mid \mathcal{Y}} \left[\nabla \breve{f}(\mathbf{x}', \mathbf{y}) \right] && (53) \\
&\approx \nabla \breve{f}(\mathbf{x}, \mathbf{y}) - \nabla \breve{f}(\bar{\mathbf{x}}(\mathbf{y}), \mathbf{y}) \,. && (54)
\end{aligned}$$

I have not tested this latter gradient scheme for BMs.

1.0.8 Expected likelihood optimisation

What happens if we want to use discriminative training, but do not know $\mathbf{y}$?

We recall from a previous section that the traditional approach for unsupervised learning, when $\mathbf{y}$ is always unknown, is to maximise the marginal likelihood, $p(\mathbf{x})$. Conversely, for supervised learning, when $\mathbf{y}$ is always known, we may optimise either $p(\mathbf{x}, \mathbf{y})$ or $p(\mathbf{y} \mid \mathbf{x})$, as we see fit (again, refer to the previous sections above). Now, for semi-supervised learning, where $\mathbf{y}$ is known for some cases but unknown for others, it is traditional to use either $p(\mathbf{x}, \mathbf{y})$ or $p(\mathbf{y} \mid \mathbf{x})$ for the cases where $\mathbf{y}$ is known, but instead to use $p(\mathbf{x})$ for the cases where $\mathbf{y}$ is unknown.

What is wrong with this traditional approach? The answer is that, for discrete distributions, we have $p(\mathbf{x}) \approx \frac{1}{|\mathcal{X}|}$, $p(\mathbf{y} \mid \mathbf{x}) \approx \frac{1}{|\mathcal{Y}|}$, and $p(\mathbf{x}, \mathbf{y}) \approx \frac{1}{|\mathcal{X}|\,|\mathcal{Y}|}$, in terms of approximate magnitudes. Thus, for cases where we have computed $p(\mathbf{x})$ in place of $p(\mathbf{x}, \mathbf{y})$, we have overestimated the joint likelihood by a factor of $|\mathcal{Y}|$. Likewise, for cases where we have computed $p(\mathbf{x})$ in place of $p(\mathbf{y} \mid \mathbf{x})$, we have overestimated the discriminative likelihood by a factor of $\frac{|\mathcal{X}|}{|\mathcal{Y}|}$. In practice, this means that traditional semi-supervised learning gives more (possibly much more) weight to unknown cases than to known cases!

The solution is to either correct the magnitude of the likelihood approximation, or else to use expected likelihoods (or, rather, expected log-likelihoods), for unsupervised and especially semi-supervised learning. Thus, if we wish to optimise the joint log-likelihood, $\ln p(\mathbf{x},\mathbf{y})$, when $\mathbf{y}$ is unknown, then instead of the traditional approximation

$$\ln p(\mathbf{x},\mathbf{y}) \;\approx\; \ln\sum_{\mathbf{y}'\in\mathcal{Y}'} p(\mathbf{x},\mathbf{y}') = \ln p(\mathbf{x})\,, \tag{55}$$

we could use the corrected version

$$\ln p(\mathbf{x},\mathbf{y}) \;\approx\; \frac{1}{|\mathcal{Y}|}\sum_{\mathbf{y}'\in\mathcal{Y}'} \ln p(\mathbf{x},\mathbf{y}')\,. \tag{56}$$

This has the correct magnitude, but essentially assumes that each $\mathbf{y}\in\mathcal{Y}$ is of equal importance.

Alternatively, we could instead use the expected value

$$\begin{aligned} \ln p(\mathbf{x},\mathbf{y}) &\approx \sum_{\mathbf{y}'\in\mathcal{Y}'} p(\mathbf{y}'\mid\mathbf{x})\,\ln p(\mathbf{x},\mathbf{y}') && (57)\\ &= \mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\ln p(\mathbf{x},\mathbf{y}')\right] \doteq L_J(\mathbf{x})\,, && (58) \end{aligned}$$

on the supposition that some values of $\mathbf{y}\in\mathcal{Y}$ are conditonally more likely than others. Interestingly, this means that $\ln p(\mathbf{x},\mathbf{y}) \approx \ln p(\mathbf{x},\bar{\mathbf{y}}(\mathbf{x}))$, via MFA, even though both terms remain intractable to compute.

The equivalent approximation to the discriminative log-likelihood is thus

$$\begin{aligned} \ln p(\mathbf{y}\mid\mathbf{x}) &\approx \sum_{\mathbf{y}'\in\mathcal{Y}'} p(\mathbf{y}'\mid\mathbf{x})\,\ln p(\mathbf{y}'\mid\mathbf{x}) && (59)\\ &= \mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\ln p(\mathbf{y}'\mid\mathbf{x})\right] \doteq L_D(\mathbf{x})\,. && (60) \end{aligned}$$

In this case, since $p(\mathbf{y}\mid\mathbf{x})$ is assumed to be tractable to compute (for small enough $|\mathcal{Y}|$), then the expectation is also tractable. Also, via MFA, we have $\ln p(\mathbf{y}\mid\mathbf{x}) \approx \ln p(\bar{\mathbf{y}}(\mathbf{x})\mid\mathbf{x})$.

To help with the gradient calculations, observe that, in general,

$$\begin{aligned} \nabla\mathbb{E}_{\mathcal{V}}\left[g(\mathbf{v})\right] &= \nabla\sum_{\mathbf{v}\in\mathcal{V}} p(\mathbf{v})\,g(\mathbf{v}) && (61)\\ &= \sum_{\mathbf{v}\in\mathcal{V}} \left\{p(\mathbf{v})\,\nabla g(\mathbf{v}) + \nabla p(\mathbf{v})\,g(\mathbf{v})\right\} && (62)\\ &= \sum_{\mathbf{v}\in\mathcal{V}} \left\{p(\mathbf{v})\,\nabla g(\mathbf{v}) + g(\mathbf{v})\,p(\mathbf{v})\,\nabla\ln p(\mathbf{v})\right\} && (63)\\ &= \mathbb{E}_{\mathcal{V}}\left[\nabla g(\mathbf{v})\right] + \mathbb{E}_{\mathcal{V}}\left[g(\mathbf{v})\,\nabla\ln p(\mathbf{v})\right]\,. && (64) \end{aligned}$$

Hence, for the expected discriminative log-likelihood we have

$$\begin{aligned} \nabla L_D(\mathbf{x}) &= \nabla\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{y}\mid\mathbf{x})\right] && (65)\\ &= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\ln p(\mathbf{y}\mid\mathbf{x})\right] + \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{y}\mid\mathbf{x})\,\nabla\ln p(\mathbf{y}\mid\mathbf{x})\right]\,. && (66) \end{aligned}$$

Now, from the previous section we have

$$\begin{aligned} \nabla\ln p(\mathbf{y}\mid\mathbf{x}) &= \nabla\breve{f}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y}')\right] && (67)\\ \Rightarrow \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\ln p(\mathbf{y}\mid\mathbf{x})\right] &= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y}')\right] \equiv 0\,. && (68) \end{aligned}$$

Note that this means we cannot simply take the expectation of the gradient, but must go further and take the gradient of the expectation, giving

$$\nabla L_D(\mathbf{x}) \;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{y}\mid\mathbf{x})\,\nabla\breve{f}(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{y}\mid\mathbf{x})\right]\,\mathbb{E}_{\mathcal{Y}'|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y}')\right] \tag{69}$$

$$\;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{y}\mid\mathbf{x})\,\nabla\breve{f}(\mathbf{x},\mathbf{y})\right] - L_D(\mathbf{x})\,\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y})\right]\,. \tag{70}$$

However, note that MFA cannot help us much here, since it results in this last difference being approximated by zero!

Now turning back to the expected joint log-likelihood, we have

$$\nabla L_J(\mathbf{x}) \;=\; \nabla\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{x},\mathbf{y})\right] \tag{71}$$

$$\;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\ln p(\mathbf{x},\mathbf{y})\right] + \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{x},\mathbf{y})\,\nabla\ln p(\mathbf{y}\mid\mathbf{x})\right]\,. \tag{72}$$

Now, from the previous section on joint likelihood optimisation we have

$$\nabla\ln p(\mathbf{x},\mathbf{y}) \;\approx\; \nabla\breve{f}(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathcal{X}'|\mathcal{Y}}\left[\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right] \tag{73}$$

$$\Rightarrow \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\ln p(\mathbf{x},\mathbf{y})\right] \;\approx\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\mathbb{E}_{\mathcal{X}'|\mathcal{Y}}\left[\mathbb{E}_{\mathcal{Y}'|\mathcal{X}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}')\right]\right]\right]\,. \tag{74}$$

Note that this is exactly the approximation for $\nabla\ln p(\mathbf{x})$, from the previous section on marginal likelihood optimisation! Hence, we could, if we wished, stop with the expectation of the gradient, namely

$$\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\nabla\ln p(\mathbf{x},\mathbf{y})\right] \;\approx\; \nabla\ln p(\mathbf{x})\,. \tag{75}$$

However, the approxpriate log-likelihood score is not $\ln p(\mathbf{x})$, for the reasons outlined at the start of this section, but instead

$$L_J(\mathbf{x}) \;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{x},\mathbf{y})\right] \tag{76}$$

$$\;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\ln p(\mathbf{x}) + \ln p(\mathbf{y}\mid\mathbf{x})\right] \tag{77}$$

$$\;=\; \ln p(\mathbf{x}) + L_D(\mathbf{x})\,. \tag{78}$$

Thus, we also see that gradient of the expectation of the joint log-likelihood is

$$\nabla L_J(\mathbf{x}) \;=\; \nabla\ln p(\mathbf{x}) + \nabla L_D(\mathbf{x})\,. \tag{79}$$

1.0.9 Hidden models

In the above derivations, we partitioned $\mathbf{v}$ into $\mathbf{x}$ and $\mathbf{y}$, where we assumed for training that $\mathbf{x}$ is always known, and $\mathbf{y}$ might or might not be known.

We now turn to the related case where $\mathbf{v}\in\mathcal{V}$ is partitioned into $\mathbf{x}\in\mathcal{X}$, $\mathbf{y}\in\mathcal{Y}$ and $\mathbf{z}\in\mathcal{Z}$, where $\mathbf{z}$ is never observed, i.e. it is latent or hidden. For convenience, we redefine $f(\mathbf{v}) = f(\breve{\mathbf{v}}(\mathbf{x},\mathbf{y},\mathbf{z})) \doteq \breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})$.

The relvant unconditional distributions are now given by

$$p(\mathbf{x},\mathbf{y},\mathbf{z}) \;=\; \frac{e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}}\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')}}\,, \tag{80}$$

$$p(\mathbf{x},\mathbf{y}) \;=\; \frac{\sum_{\mathbf{z}\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}}\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')}}\,, \tag{81}$$

$$p(\mathbf{x}) \;=\; \frac{\sum_{\mathbf{y}\in\mathcal{Y}}\sum_{\mathbf{z}\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}}\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')}}\,. \tag{82}$$

Similarly, the relevant conditional distributions are

$$p(\mathbf{x},\mathbf{y}\mid\mathbf{z}) \;=\; \frac{e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}} e^{\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z})}}\,, \tag{83}$$

$$p(\mathbf{z}\mid\mathbf{x},\mathbf{y}) \;=\; \frac{e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z}')}}\,. \tag{84}$$

The forward and backward predictive distributions are

$$p(\mathbf{y}\mid\mathbf{x}) \;=\; \frac{\sum_{\mathbf{z}\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{y}'\in\mathcal{Y}}\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y}',\mathbf{z}')}}\,, \tag{85}$$

$$p(\mathbf{x}\mid\mathbf{y}) \;=\; \frac{\sum_{\mathbf{z}\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})}}{\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x}',\mathbf{y},\mathbf{z}')}}\,. \tag{86}$$

1.0.10 Joint likelihood optimisation

From the above definition of $p(\mathbf{x},\mathbf{y})$, we have

$$\ln p(\mathbf{x},\mathbf{y}) \;=\; \ln\sum_{\mathbf{z}\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})} - \ln\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}}\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')} \tag{87}$$

$$\Rightarrow \nabla\ln p(\mathbf{x},\mathbf{y}) \;=\; \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})\right] - \mathbb{E}_{\mathcal{X}',\mathcal{Y}',\mathcal{Z}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')\right]\,. \tag{88}$$

Examination of the various gradient approximations derived in earlier sections suggests yet another pattern, namely that the inner conditional expectation on the right-hand side of the difference should match the conditional expectation on the left-hand side (with added primes). Thus, we choose the CEA expansion

$$\nabla\ln p(\mathbf{x},\mathbf{y}) \;=\; \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})\right] - \mathbb{E}_{\mathcal{Z}}\left[\mathbb{E}_{\mathcal{X}',\mathcal{Y}'|\mathcal{Z}}\left[\mathbb{E}_{\mathcal{Z}'|\mathcal{X}',\mathcal{Y}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')\right]\right]\right] \tag{89}$$

$$\approx\; \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})\right] - \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}}\left[\mathbb{E}_{\mathcal{X}',\mathcal{Y}'|\mathcal{Z}}\left[\mathbb{E}_{\mathcal{Z}'|\mathcal{X}',\mathcal{Y}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')\right]\right]\right]\,. \tag{90}$$

We lack sufficient information about the specific model to be able to approximate the middle expectation $\mathbb{E}_{\mathcal{X}',\mathcal{Y}'|\mathcal{Z}}$ via MFA. However, for the inner and outer expectations, it is clear that we need to define

$$\bar{\mathbf{z}}(\mathbf{x},\mathbf{y}) \;\doteq\; \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}}[\mathbf{z}]\,. \tag{91}$$

Consequently, we at least know that

$$\nabla \ln p(\mathbf{x},\mathbf{y}) \;\approx\; \nabla \breve{f}(\mathbf{x},\mathbf{y},\bar{\mathbf{z}}) - \nabla \breve{f}(\bar{\mathbf{x}}',\bar{\mathbf{y}}',\bar{\mathbf{z}}')\,, \tag{92}$$

where $\bar{\mathbf{z}} = \bar{\mathbf{z}}(\mathbf{x},\mathbf{y})$ and $\bar{\mathbf{z}}' = \bar{\mathbf{z}}(\mathbf{x}',\mathbf{y}')$.

In order to compute the log-likelihood score, observe that

$$p(\mathbf{x},\mathbf{y}) \;=\; \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{y},\mathbf{z}) \tag{93}$$

$$=\; \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{y}\mid\mathbf{z})\,p(\mathbf{z}) \tag{94}$$

$$=\; \mathbb{E}_{\mathcal{Z}}[p(\mathbf{x},\mathbf{y}\mid\mathbf{z})] \tag{95}$$

$$\approx\; \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}}[p(\mathbf{x},\mathbf{y}\mid\mathbf{z})] \tag{96}$$

$$\approx\; p(\mathbf{x},\mathbf{y}\mid\bar{z}(\mathbf{x},\mathbf{y}))\,, \tag{97}$$

via CEA and MFA.

1.0.11 Marginal likelihood optimisation

From the definition of $p(\mathbf{x})$ in a previous section, we have

$$\ln p(\mathbf{x}) \;=\; \ln \sum_{\mathbf{y}\in\mathcal{Y}}\sum_{\mathbf{z}\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})} - \ln \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\mathbf{y}'\in\mathcal{Y}}\sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')} \tag{98}$$

$$\Rightarrow \nabla \ln p(\mathbf{x}) \;=\; \mathbb{E}_{\mathcal{Y},\mathcal{Z}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})\right] - \mathbb{E}_{\mathcal{X}',\mathcal{Y}',\mathcal{Z}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')\right] \tag{99}$$

$$\approx\; \mathbb{E}_{\mathcal{Y},\mathcal{Z}|\mathcal{X}}\left[\nabla\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})\right] - \mathbb{E}_{\mathcal{Y},\mathcal{Z}|\mathcal{X}}\left[\mathbb{E}_{\mathcal{X}'|\mathcal{Y},\mathcal{Z}}\left[\mathbb{E}_{\mathcal{Y}',\mathcal{Z}'|\mathcal{X}'}\left[\nabla\breve{f}(\mathbf{x}',\mathbf{y}',\mathbf{z}')\right]\right]\right]\,, \tag{100}$$

via CEA. We further note that

$$\mathbb{E}_{\mathcal{Y},\mathcal{Z}|\mathcal{X}}[\cdot] \;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}}[\cdot]\right] \tag{101}$$

$$\Rightarrow \mathbb{E}_{\mathcal{Y},\mathcal{Z}|\mathcal{X}}[\mathbf{y}] \;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\mathbf{y}] = \bar{\mathbf{y}}(\mathbf{x})\,, \tag{102}$$

$$\mathbb{E}_{\mathcal{Y},\mathcal{Z}|\mathcal{X}}[\mathbf{z}] \;=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\bar{\mathbf{z}}(\mathbf{x},\mathbf{y})] \approx \bar{\mathbf{z}}(\mathbf{x},\bar{\mathbf{y}}(\mathbf{x}))\,, \tag{103}$$

via MFA. Consequently, we have

$$\nabla \ln p(\mathbf{x}) \;\approx\; \nabla\breve{f}(\mathbf{x},\bar{\mathbf{y}},\bar{\mathbf{z}}) - \nabla\breve{f}(\bar{\mathbf{x}}',\bar{\mathbf{y}}',\bar{\mathbf{z}}')\,, \tag{104}$$

where $\bar{\mathbf{y}} = \bar{\mathbf{y}}(\mathbf{x})$, $\bar{\mathbf{z}} = \bar{\mathbf{z}}(\mathbf{x},\bar{\mathbf{y}})$, $\bar{\mathbf{y}}' = \bar{\mathbf{y}}(\bar{\mathbf{x}}')$, and $\bar{\mathbf{z}}' = \bar{\mathbf{z}}(\bar{\mathbf{x}}',\bar{\mathbf{y}}')$. The form that $\bar{\mathbf{x}}'$ takes depends upon $\mathbb{E}_{\mathcal{X}'|\mathcal{Y},\mathcal{Z}}$.

In order to compute the log-likelihood score, observe that

$$p(\mathbf{x}) \;=\; \sum_{\mathbf{y}\in\mathcal{Y}}\sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{y},\mathbf{z}) \tag{105}$$

$$=\; \sum_{\mathbf{y}\in\mathcal{Y}}\sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x}\mid\mathbf{y},\mathbf{z})\,p(\mathbf{y},\mathbf{z}) \tag{106}$$

$$=\; \mathbb{E}_{\mathcal{Y},\mathcal{Z}}[p(\mathbf{x}\mid\mathbf{y},\mathbf{z})] \tag{107}$$

$$\approx\; \mathbb{E}_{\mathcal{Y},\mathcal{Z}|\mathcal{X}}[p(\mathbf{x}\mid\mathbf{y},\mathbf{z})] \tag{108}$$

$$=\; \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\mathbb{E}_{\mathcal{Z}|\mathcal{Y},\mathcal{X}}[p(\mathbf{x}\mid\mathbf{y},\mathbf{z})]\right] \tag{109}$$

$$\approx\; p(\mathbf{x}\mid\bar{\mathbf{y}}(\mathbf{x}),\bar{z}(\mathbf{x},\bar{\mathbf{y}}(\mathbf{x})))\,, \tag{110}$$

via CEA and MFA.

1.0.12 Conditional likelihood optimisation

From the definition of $p(\mathbf{y} \mid \mathbf{x})$ in an earlier section, we have

$$\begin{aligned}
\ln p(\mathbf{y} \mid \mathbf{x}) &= \ln \sum_{\mathbf{z}\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})} - \ln \sum_{\mathbf{y}'\in\mathcal{Y}} \sum_{\mathbf{z}'\in\mathcal{Z}} e^{\breve{f}(\mathbf{x},\mathbf{y}',\mathbf{z}')} && (111) \\
\Rightarrow \nabla \ln p(\mathbf{y} \mid \mathbf{x}) &= \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}} \left[\nabla \breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})\right] - \mathbb{E}_{\mathcal{Y}',\mathcal{Z}'|\mathcal{X}} \left[\nabla \breve{f}(\mathbf{x},\mathbf{y}',\mathbf{z}')\right] && (112) \\
&= \mathbb{E}_{\mathcal{Z}|\mathcal{X},\mathcal{Y}} \left[\nabla \breve{f}(\mathbf{x},\mathbf{y},\mathbf{z})\right] - \mathbb{E}_{\mathcal{Y}'|\mathcal{X}} \left[\mathbb{E}_{\mathcal{Z}'|\mathcal{X},\mathcal{Y}'} \left[\nabla \breve{f}(\mathbf{x},\mathbf{y}',\mathbf{z}')\right]\right] && (113) \\
&\approx \nabla \breve{f}(\mathbf{x},\mathbf{y},\bar{\mathbf{z}}(\mathbf{x},\mathbf{y})) - \nabla \breve{f}(\mathbf{x},\bar{\mathbf{y}}(\mathbf{x}),\bar{\mathbf{z}}(\mathbf{x},\bar{\mathbf{y}}(\mathbf{x})))\,, && (114)
\end{aligned}$$

via MFA.

We can therefore write this as

$$\nabla \ln p(\mathbf{y} \mid \mathbf{x}) \;\approx\; \nabla \breve{f}(\mathbf{x},\mathbf{y},\bar{\mathbf{z}}) - \nabla \breve{f}(\mathbf{x},\bar{\mathbf{y}}',\bar{\mathbf{z}}')\,, \qquad (115)$$

in contrast to the above gradient of the joint log-likelihood, namely

$$\nabla \ln p(\mathbf{x},\mathbf{y}) \;\approx\; \nabla \breve{f}(\mathbf{x},\mathbf{y},\bar{\mathbf{z}}) - \nabla \breve{f}(\bar{\mathbf{x}}',\bar{\mathbf{y}}',\bar{\mathbf{z}}')\,. \qquad (116)$$

Thus, when directly optimising the conditional (or discriminative) log-likelhood, we do not reconstruct the input $\mathbf{x}$ via $\bar{\mathbf{x}}'$. In practice, this means that there are some parameters (i.e. those linked entirely to $\mathbf{x}$) that cannot be directly estimated (but might be indirectly estimated via some hybrid gradient scheme).