Variational inference: Suppose we have a generative model of some data $d$ characterized by some hidden variables $\alpha$ and parameters $\theta$. The loglikelihood function can be written as follows:

$$\log p(d|\theta) = \log \left[ \sum_{\alpha} p(d, \alpha|\theta) \right] = \log \left[ \sum_{\alpha} p(d|\alpha)p(\alpha|\theta) \right]$$

Equivalently, we can define an energy function as follows and express everything in terms of it:

$$E_\alpha = -\log p(d, \alpha|\theta)$$

where we suppressed the dependence of $E_\alpha$ on $d$ and $\theta$ for simplicity of notation. In terms of the energy function, we can write down the posterior probability of $\alpha$ given $d$ as follows:

$$P_\alpha \equiv p(\alpha|d, \theta) = \frac{p(\alpha, d|\theta)}{\sum_{\alpha'} p(\alpha', d|\theta)} = \exp(-E_\alpha) \sum_{\alpha'} \exp(-E_{\alpha'})$$

This is known as the Boltzmann distribution. The denominator of the last expression is called the partition function and it corresponds to the marginal likelihood of the data. We shall denote it by $Z$, as is customary. It is arguably the single most important quantity in statistical mechanics. Now, we shall try to express $\log Z$ in terms of $E_\alpha$ and $P_\alpha$. From the Boltzmann distribution, we see that $Z = \exp(-E_\alpha) P_\alpha$. Then:

$$\log Z = \log \left[ \frac{\exp(-E_\alpha)}{P_\alpha} \right] = -E_\alpha - \log P_\alpha$$

and this holds for all $\alpha$. It also holds if we take the expected value of both sides in Equation 4 with respect to $P_\alpha$ (note that $\log Z$ in fact does not depend on $\alpha$ so this is a dummy operation for the left hand side):

$$\log Z = \langle -E_\alpha - \log P_\alpha \rangle_{P_\alpha}$$

In fact, since taking the expected value of the left hand-side is a dummy operation, we can do this for any arbitrary distribution $Q_\alpha$ over $\alpha$. Suppose, in particular, that $P_\alpha$ is an intractable distribution and we wish to approximate it with a simpler distribution $Q_\alpha$. Then we have:

$$\log Z = \langle -E_\alpha - \log P_\alpha \rangle_{Q_\alpha} = -\sum_\alpha Q_\alpha E_\alpha - \sum_\alpha Q_\alpha \log Q_\alpha + \sum_\alpha Q_\alpha \log \frac{Q_\alpha}{P_\alpha}$$

Now this equation gives us a prescription for obtaining the $Q$ that approximates $P$ best. Note that the left hand side, $\log Z$, is independent of $Q$, it is just the marginal loglikelihood of the data. The second term on the right hand side is $\text{KL}[Q, P]$, the KL-divergence between $Q$ and $P$ and we know that it has to be non-negative. Ideally, we want to make $\text{KL}[Q, P]$ as small as possible. Equation 6 suggests that we can minimize $\text{KL}[Q, P]$ by maximizing $-\mathcal{F}(E, Q)$ ($\mathcal{F}$ is called the Helmholtz free energy). This works because $\text{KL}[Q, P]$ and $-\mathcal{F}(E, Q)$ add up to a constant (i.e. $\log Z$), hence by maximizing $-\mathcal{F}(E, Q)$, we are necessarily minimizing $\text{KL}[Q, P]$.

The Helmholtz machine: The Helmholtz machine is a particular instantiation of the variational inference frame-
work described above. It is a neural network that has an input layer and a number of hidden layers stacked on top of each other. The consecutive layers are connected through both feedforward and feedback connections. The feedforward connections are also known as the recognition weights and the feedback connections are known as the generative weights for reasons that will be clear shortly.

In the previous subsection, we found that we need to calculate the Helmholtz free energy $\mathcal{F}(E, Q)$ where $E$ is the energy function and $Q$ is the approximation to the posterior. To make this calculation tractable, we are going to assume that $Q$ is a separable distribution in each layer (we shall make a similar assumption for $E$ as well). This means that given the activities of all units in layer $l$, the units in layer $l+1$ are conditionally independent. More specifically, $Q_\alpha$ takes the following form:

$$Q_\alpha = \prod_{l>1} \prod_j [q_j^l(\phi, s^{l-1})]^{s_j^l} [1 - q_j^l(\phi, s^{l-1})]^{1 - s_j^l}$$  \hspace{1cm} (7)$$

Here, $s_j^l$ denotes the binary stochastic activity of the $j$-th unit in the $l$-th layer, $s^l$ is the vector of activities of all units in the $l$-th layer and $\phi_{i,j}^l$ is the probability of being active for the $j$-th neuron in the $l$-th layer and is given by:

$$q_j^l(\phi, s^{l-1}) = \sigma(\sum_i s_i^{l-1} \phi_{i,j}^{l-1,l})$$  \hspace{1cm} (8)$$

where $\sigma(\cdot)$ is the sigmoid function and $\phi_{i,j}^{l-1,l}$ is the feedforward (recognition) weight between the $i$-th unit in layer $l-1$ and the $j$-th unit in layer $l$. In practice, Dayan et al. (1995) use an approximation where the stochastic activities $s_j^l$ are replaced by their means $\bar{q}_j^l$, as in mean field models:

$$q_j^l(\phi, \bar{q}^{l-1}) = \sigma(\sum_i \bar{q}_i^{l-1} \phi_{i,j}^{l-1,l})$$  \hspace{1cm} (9)$$

We make a similar separability assumption for the joint distribution $p(d, \alpha|\theta)$ that determines the energy function $E_\alpha$ (Equation 2):

$$p(d, \alpha|\theta) = \prod_{l>1} \prod_j [p_j^l(\theta, s^{l-1})]^{s_j^l} [1 - p_j^l(\theta, s^{l-1})]^{1 - s_j^l}$$  \hspace{1cm} (10)$$

where $p_j^l$ is the activation probability of the $j$-th unit in layer $l$ in the generative model for which they use a mean-field type approximation similar to the one used for $q_j^l$ above:

$$p_j^l(\theta, q^{l+1}) = \left(1 - \frac{1}{1 + \sum_k \theta_{k,j}^{l+1} q_j^{l+1}}\right) \left(1 - \prod_k \left[1 - q_k^{l+1} \frac{\theta_{k,j}^{l+1}}{1 + \theta_{k,j}^{l+1}}\right]\right)$$  \hspace{1cm} (11)$$

The reason they don’t use the obvious analogue of Equation 9 for $p_j^l$ is that it turns out not to work well in practice due to local minima. Equation 11 is really just a hack to escape local minima.

The first layer activations $q^1$ constitute the data $d$ in the model. With this factorized approximation to $p(d, \alpha|\theta)$, $-\mathcal{F}(E, Q)$ can be written as:

$$-\mathcal{F}(E, Q) = \langle -E_\alpha - \log Q_\alpha \rangle_{Q_\alpha}$$

$$\propto \sum_d \sum_l \sum_j q_j^l \log \frac{q_j^l}{p_j^l} + (1 - q_j^l) \log \frac{1 - q_j^l}{1 - p_j^l}$$  \hspace{1cm} (13)$$

where $q_j^l$ and $p_j^l$ are given by Equations 9 and 11 respectively and the sum over $d$ represents a sum (or average) over the data. The derivatives of $\mathcal{F}(E, Q)$ with respect to the recognition and the generative weights, $\phi$ and $\theta$, can be easily calculated using the chain rule and they are given in the Appendix of Dayan et al. (1995).

References