1 Basics

1.1 Configuration Space

Let $z \in \mathbb{R}^E$ represent the embeding vector, $m = 1, \ldots, M$ is the categorical label, and $q_m(z, \theta) := \mathrm{softmax}_m(f(z, \theta))$ with $f(\cdot, \theta)$ a neural network parameterized by θ. Given (z, θ), we have

$$\frac{\partial}{\partial \theta} \ln q_m = \frac{\partial f_m}{\partial \theta} - \sum_\alpha q_\alpha \frac{\partial f_\alpha}{\partial \theta}. \tag{1}$$

Consider $f_\alpha(z, \theta) = \sum_\beta U_{\alpha\beta}\, \sigma(\sum_\gamma W_{\beta\gamma} z_\gamma + b_\beta) + c_\alpha$, where σ represents the ReLU function, with $\sigma^{(n)}(0) = 1$ only when $n = 1$ otherwise zero. So, when $\theta = 0$ (i.e. $U, c, W, b = 0$), all the non-vanishing terms are[1]

$$\frac{\partial f_\alpha}{\partial c_\alpha}(z, 0) = 1; \tag{2}$$

$$\frac{\partial^2 f_\alpha}{\partial U_{\alpha\beta} \partial W_{\beta\gamma}}(z, 0) = z_\gamma; \tag{3}$$

and

$$\frac{\partial^2 f_\alpha}{\partial U_{\alpha\beta} \partial b_\beta}(z, 0) = 1. \tag{4}$$

1.2 Data and Action

Given the distribution of real world data p, the relative entropy between p and q is

$$H[p, q] = \sum_{z,m} p(z, m) \ln p(z, m) - \sum_{z,m} p(z, m) \ln q_m(z, \theta).$$

1. The non-vanishing terms of the first derivative are

$$\begin{aligned}
\frac{\partial f_\alpha}{\partial U_{\alpha\beta}}(z, \theta) &= \sigma\left(\sum_\gamma W_{\beta\gamma} z_\gamma + b_\beta\right); \\
\frac{\partial f_\alpha}{\partial c_\alpha}(z, \theta) &= 1; \\
\frac{\partial f_\alpha}{\partial W_{\beta\gamma}}(z, \theta) &= U_{\alpha\beta}\, \sigma'\left(\sum_{\gamma'} W_{\beta\gamma'} z_{\gamma'} + b_\beta\right) z_\gamma; \\
\frac{\partial f_\alpha}{\partial b_\beta}(z, \theta) &= U_{\alpha\beta}\, \sigma'\left(\sum_{\gamma'} W_{\beta\gamma'} z_{\gamma'} + b_\beta\right).
\end{aligned}$$

Thus, the non-vanishing terms of the second derivative are

$$\begin{aligned}
\frac{\partial^2 f_\alpha}{\partial U_{\alpha\beta} \partial W_{\beta\gamma}}(z, \theta) &= \sigma'\left(\sum_{\gamma'} W_{\beta\gamma'} z_{\gamma'} + b_\beta\right) z_\gamma; \\
\frac{\partial^2 f_\alpha}{\partial U_{\alpha\beta} \partial b_\beta}(z, \theta) &= \sigma'\left(\sum_{\gamma'} W_{\beta\gamma'} z_{\gamma'} + b_\beta\right); \\
\frac{\partial^2 f_\alpha}{\partial W_{\beta\gamma} \partial W_{\beta\gamma'}}(z, \theta) &= U_{\alpha\beta}\, \sigma''\left(\sum_{\gamma''} W_{\beta\gamma''} z_{\gamma''} + b_\beta\right) z_\gamma z_{\gamma'}; \\
\frac{\partial^2 f_\alpha}{\partial W_{\beta\gamma} \partial b_\beta}(z, \theta) &= U_{\alpha\beta}\, \sigma''\left(\sum_{\gamma'} W_{\beta\gamma'} z_{\gamma'} + b_\beta\right) z_\gamma; \\
\frac{\partial^2 f_\alpha}{\partial b_\beta \partial b_\beta}(z, \theta) &= U_{\alpha\beta}\, \sigma''\left(\sum_{\gamma'} W_{\beta\gamma'} z_{\gamma'} + b_\beta\right).
\end{aligned}$$

For higher derivatives, $\sigma^{(1)}$ is absent.

The first term is θ-independent. Thus, the action of θ shall be the second term, that is

$$S(\theta) := -\sum_{z,m} p(z,m) \ln q_m(z,\theta). \tag{5}$$

Assume that $p(m) := \sum_z p(z,m) = 1/M$ for all $m = 1,\ldots,M$, meaning that the data have been properly balanced.

2 Taylor Expansion

Now, we are to Taylor expand $S(\theta)$ at $\theta = 0$. Denote the expansion by

$$S(\theta) =: S_0 + S_1(\theta) + \cdots, \tag{6}$$

where $S_n(\theta) \sim \theta^n$, and $S_0 := S(0)$ is θ-independent.

2.1 Zeroth-Order

When $\theta = 0$ (i.e. $U, c, W, b = 0$), we have $f_\alpha(z,0) = 0$, thus $q_\alpha(z,0) = \text{softmax}_\alpha(f(z,0)) = 1/M$ for all $\alpha = 1,\ldots,M$. So,

$$S_0 = \ln M. \tag{7}$$

2.2 First-Order

Plugging in equation 1, we have

$$\frac{\partial S}{\partial \theta} = \sum_{z,m} p(z,m) \left[\sum_\alpha q_\alpha \frac{\partial f_\alpha}{\partial \theta} - \frac{\partial f_m}{\partial \theta} \right].$$

At $\theta = 0$, all terms are vanishing.[2] So,

$$S_1(\theta) = 0. \tag{8}$$

2.3 Second-Order

Taking derivative on $\partial S / \partial \theta$ and plugging in equation 1, we arrive at

$$\frac{\partial^2 S}{\partial \theta \partial \theta'} = \sum_{z,m} p(z,m) \left[\sum_\alpha q_\alpha \left(\frac{\partial f_\alpha}{\partial \theta} \frac{\partial f_\alpha}{\partial \theta'} + \frac{\partial^2 f_\alpha}{\partial \theta \partial \theta'} \right) - \sum_{\alpha,\beta} q_\alpha q_\beta \frac{\partial f_\alpha}{\partial \theta} \frac{\partial f_\beta}{\partial \theta'} - \frac{\partial^2 f_m}{\partial \theta \partial \theta'} \right].$$

At $\theta = 0$, we have

$$\frac{\partial^2 S}{\partial c_\alpha \partial c_\beta}(0) = \frac{\delta_{\alpha\beta}}{M} - \frac{1}{M^2},$$

2. By equations 2, 3, and 4, at the first order, the only term that is not apparently zero is derivative on c. But,

$$\begin{aligned}
\frac{\partial S}{\partial c_\alpha}(0) &= \sum_{z,m} p(z,m) \left[\sum_{\alpha'} q_{\alpha'} \frac{\partial f_{\alpha'}}{\partial c_\alpha} - \frac{\partial f_m}{\partial c_\alpha} \right] \\
\left\{ \frac{\partial f}{\partial c} = \cdots \right\} &= \sum_{z,m} p(z,m) \left[\frac{1}{M} \sum_{\alpha'} \delta_{\alpha\alpha'} - \delta_{m\alpha} \right] \\
\left\{ p(m) = \frac{1}{M} \right\} &= \frac{1}{M} \sum_{\alpha'} \delta_{\alpha\alpha'} - \frac{1}{M} \sum_m \delta_{m\alpha} \\
&= 0.
\end{aligned}$$

and

$$\frac{\partial^2 S}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}}(0) = \frac{1}{M}(\mathbb{E}_{z\sim p(z)}[z_\gamma] - \mathbb{E}_{z\sim p(z|\alpha)}[z_\gamma]) =: J_{\alpha\gamma}.$$

Other terms are all vanishing.[3] So,

$$S_2(\theta) = \frac{1}{2}\sum_{\alpha,\beta}\left(\frac{\delta_{\alpha\beta}}{M} - \frac{1}{M^2}\right)c_\alpha c_\beta + \frac{1}{2}\sum_{\alpha,\gamma} J_{\alpha\gamma}\sum_\beta U_{\alpha\beta}W_{\beta\gamma}. \tag{9}$$

By numerical computation, we find that the matrix $\delta/M - 1/M^2$ is non-negative definite. The term $\sum_\beta U_{\alpha\beta}W_{\beta\gamma}$ can be seen as a "propagation" from the γ-neuron to the α-neuron, weighted by $J_{\alpha\gamma}$.

2.4 Third-Order (TODO)

Taking derivative on $\partial^2 S/(\partial\theta\partial\theta')$ and plugging in equation 1, we have

$$\begin{aligned}
\frac{\partial^3 S}{\partial\theta\partial\theta'\partial\theta''} = &\sum_{z,m} p(z,m)\sum_\alpha q_\alpha\left[\frac{\partial^3 f_\alpha}{\partial\theta\partial\theta'\partial\theta''} + \frac{\partial f_\alpha}{\partial\theta}\frac{\partial f_\alpha}{\partial\theta'}\frac{\partial f_\alpha}{\partial\theta''} + \frac{\partial f_\alpha}{\partial\theta}\frac{\partial^2 f_\alpha}{\partial\theta'\partial\theta''} + \frac{\partial f_\alpha}{\partial\theta''}\frac{\partial^2 f_\alpha}{\partial\theta\partial\theta'} + \frac{\partial f_\alpha}{\partial\theta'}\frac{\partial^2 f_\alpha}{\partial\theta''\partial\theta}\right]\\
&-\sum_{z,m} p(z,m)\sum_{\alpha,\beta} q_\alpha q_\beta\left[\frac{\partial^2 f_\alpha}{\partial\theta\partial\theta'}\frac{\partial f_\beta}{\partial\theta''} + \frac{\partial^2 f_\alpha}{\partial\theta\partial\theta''}\frac{\partial f_\beta}{\partial\theta'} + \frac{\partial f_\alpha}{\partial\theta}\frac{\partial^2 f_\beta}{\partial\theta'\partial\theta''} + \frac{\partial f_\alpha}{\partial\theta}\frac{\partial f_\alpha}{\partial\theta'}\frac{\partial f_\beta}{\partial\theta''} + \frac{\partial f_\alpha}{\partial\theta}\frac{\partial f_\alpha}{\partial\theta''}\frac{\partial f_\beta}{\partial\theta'} + \frac{\partial f_\alpha}{\partial\theta}\frac{\partial f_\beta}{\partial\theta'}\frac{\partial f_\beta}{\partial\theta''}\right]\\
&+2\sum_{z,m} p(z,m)\sum_{\alpha,\beta,\gamma} q_\alpha q_\beta q_\gamma\frac{\partial f_\alpha}{\partial\theta}\frac{\partial f_\beta}{\partial\theta'}\frac{\partial f_\gamma}{\partial\theta''}\\
&-\sum_{z,m} p(z,m)\frac{\partial^3 f_m}{\partial\theta\partial\theta'\partial\theta''}
\end{aligned}$$

3. Plugging in equation 2, we have

$$\begin{aligned}
\frac{\partial^2 S}{\partial c_\alpha\partial c_\beta}(0) &= \sum_{z,m} p(z,m)\left[\sum_\gamma q_\gamma\frac{\partial f_\gamma}{\partial c_\alpha}\frac{\partial f_\gamma}{\partial c_\beta} - \sum_{\gamma,\gamma'} q_\gamma q_{\gamma'}\frac{\partial f_\gamma}{\partial c_\alpha}\frac{\partial f_{\gamma'}}{\partial c_\beta}\right]\\
\left\{q_\alpha = \frac{1}{M}, \frac{\partial f}{\partial c} = \cdots\right\} &= \sum_{z,m} p(z,m)\left[\frac{1}{M}\delta_{\alpha\beta} - \frac{1}{M^2}\right]\\
\left\{\sum_{z,m} p(z,m) = 1\right\} &= \frac{\delta_{\alpha\beta}}{M} - \frac{1}{M^2}.
\end{aligned}$$

Plugging in equation 3, we have

$$\begin{aligned}
\frac{\partial^2 S}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}}(0) &= \sum_{z,m} p(z,m)\left[\sum_{\alpha'} q_{\alpha'}\frac{\partial^2 f_{\alpha'}}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}} - \frac{\partial^2 f_m}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}}\right]\\
\left\{q_\alpha = \frac{1}{M}, \frac{\partial^2 f}{\partial U\partial W} = \cdots\right\} &= \sum_{z,m} p(z,m)\left[\frac{1}{M}z_\gamma - \delta_{m\alpha}z_\gamma\right]\\
&= \frac{1}{M}\sum_z p(z)\,z_\gamma - \sum_z p(z,\alpha)\,z_\gamma\\
\{p(z,\alpha) = p(\alpha)\,p(z|\alpha)\} &= \frac{1}{M}\sum_z p(z)\,z_\gamma - p(\alpha)\sum_z p(z|\alpha)\,z_\gamma\\
\left\{p(\alpha) = \frac{1}{M}\right\} &= \frac{1}{M}\left(\sum_z p(z)\,z_\gamma - \sum_z p(z|\alpha)\,z_\gamma\right)\\
&= \frac{1}{M}(\mathbb{E}_{z\sim p(z)}[z_\gamma] - \mathbb{E}_{z\sim p(z|\alpha)}[z_\gamma]).
\end{aligned}$$

Plugging in equation 4, we have

$$\begin{aligned}
\frac{\partial^2 S}{\partial U_{\alpha\beta}\partial b_\beta}(0) &= \sum_{z,m} p(z,m)\left[\sum_{\alpha'} q_{\alpha'}\frac{\partial^2 f_{\alpha'}}{\partial U_{\alpha\beta}\partial b_\beta} - \frac{\partial^2 f_m}{\partial U_{\alpha\beta}\partial b_\beta}\right]\\
\left\{q_\alpha = \frac{1}{M}, \frac{\partial^2 f}{\partial U\partial b} = \cdots\right\} &= \sum_{z,m} p(z,m)\left[\frac{1}{M}\sum_{\alpha'}\delta_{\alpha\alpha'} - \delta_{m\alpha}\right]\\
\left\{p(m) = \frac{1}{M}\right\} &= \frac{1}{M}\sum_{\alpha'}\delta_{\alpha\alpha'} - \frac{1}{M}\sum_m \delta_{m\alpha}\\
&= 0.
\end{aligned}$$

Plugging in equation 2, 3, and 4, the terms that are not apparently zero come to be

$$
\begin{aligned}
\frac{\partial^3 S}{\partial c_\alpha \partial c_\beta \partial c_\gamma}(0) = &\sum_{z,m} p(z,m) \sum_{\alpha'} q_{\alpha'} \left[\frac{\partial f_{\alpha'}}{\partial c_\alpha} \frac{\partial f_{\alpha'}}{\partial c_\beta} \frac{\partial f_{\alpha'}}{\partial c_\gamma} \right] \\
&- \sum_{z,m} p(z,m) \sum_{\alpha',\beta'} q_{\alpha'} q_{\beta'} \left[\frac{\partial f_{\alpha'}}{\partial c_\alpha} \frac{\partial f_{\alpha'}}{\partial c_\beta} \frac{\partial f_{\beta'}}{\partial c_\gamma} + \frac{\partial f_{\alpha'}}{\partial c_\alpha} \frac{\partial f_{\beta'}}{\partial c_\beta} \frac{\partial f_{\alpha'}}{\partial c_\gamma} + \frac{\partial f_{\alpha'}}{\partial c_\alpha} \frac{\partial f_{\beta'}}{\partial c_\beta} \frac{\partial f_{\beta'}}{\partial c_\gamma} \right] \\
&+ 2 \sum_{z,m} p(z,m) \sum_{\alpha',\beta',\gamma'} q_{\alpha'} q_{\beta'} q_{\gamma'} \frac{\partial f_{\alpha'}}{\partial c_\alpha} \frac{\partial f_{\beta'}}{\partial c_\beta} \frac{\partial f_{\gamma'}}{\partial c_\gamma} \\
= &\frac{\delta_{\alpha\beta\gamma}}{M} - \frac{1}{M^2} \sum_{\alpha',\beta'} \left[\delta_{\alpha\alpha'} \delta_{\beta\alpha'} \delta_{\gamma\beta'} + \delta_{\alpha\alpha'} \delta_{\beta\beta'} \delta_{\gamma\alpha'} + \delta_{\alpha\alpha'} \delta_{\beta\beta'} \delta_{\gamma\beta'} \right] + \frac{2}{M^3} \sum_{\alpha',\beta',\gamma'} \delta_{\alpha\alpha'} \delta_{\beta\beta'} \delta_{\gamma\gamma'} \\
= &\frac{\delta_{\alpha\beta\gamma}}{M} - \frac{\delta_{\alpha\beta} + \delta_{\alpha\gamma} + \delta_{\beta\gamma}}{M^2} + \frac{2}{M^3}; \\
\frac{\partial^3 S}{\partial U_{\alpha\beta} \partial W_{\beta\gamma} \partial c_\delta}(0) = &? \\
\frac{\partial^3 S}{\partial U_{\alpha\beta} \partial b_\beta \partial c_\gamma}(0) = &?
\end{aligned}
$$

So,

$$
S_3(\theta) = \sum_{\alpha,\beta,\gamma} \left(\frac{\delta_{\alpha\beta\gamma}}{6M} - \frac{\delta_{\alpha\beta} + \delta_{\alpha\gamma} + \delta_{\beta\gamma}}{6M^2} + \frac{1}{3M^3} \right) c_\alpha c_\beta c_\gamma + ? \tag{10}
$$