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 $q_m(z, \theta) := \text{softmax}_m(f(z, \theta))$ [1] 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

$$f_\alpha(z, \theta) = \sum_\beta U_{\alpha\beta}\, \sigma\left(\sum_\gamma W_{\beta\gamma} z_\gamma + b_\beta\right),$$

where σ represents the SiLU activation, that is, $\sigma(x) = x / (1 + e^{-x})$. It is a smooth version of ReLU activation. Given the "hidden dimension" H, we have $U \in \mathbb{R}^{M \times H}$, $c \in \mathbb{R}^M$, $W \in \mathbb{R}^{H \times E}$, and $b \in \mathbb{R}^H$.

1.2 Data and Action

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

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

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

$$S(\theta) := -\sum_{z,m} p(z, m) \ln q_m(z, \theta). \tag{2}$$

This action has the minimum $S(\theta_\star) = H[p] := -\sum_{z,m} p(z, m) \ln p(z, m)$, where $q_m(z, \theta_\star) = 1$ for each z.

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

2 Taylor Expansion of Action

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

$$S(\theta) =: S_0 + S_1(\theta) + \cdots, \tag{3}$$

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,

$$S_0 = \ln M. \tag{4}$$

1. Softmax function is defined by softmax: $\mathbb{R}^n \to \mathbb{R}^n$ with

$$\text{softmax}_\alpha(x) := \frac{\exp(x_\alpha)}{\sum_\beta \exp(x_\beta)}.$$

2.2 First Order

Plugging in equation 1, we have

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

To calculate $(\partial S/\partial \theta)(0)$, we have to calculate $(\partial f/\partial \theta)(z,0)$. Replacing θ by U, W, and b respectively, we have the non-vanishing terms

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

Setting $\theta=0$, nothing is left. Thus,

$$S_1(\theta)=0. \tag{5}$$

2.3 Second Order

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

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

To calculate $(\partial^2 S/\partial \theta \partial \theta')(0)$, we have to calculate $(\partial^2 f/\partial \theta \partial \theta')(z,0)$. We have the non-vanishing terms

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

Since $\sigma(0)=0$, $\sigma'(0)=1/2$, we have, in addition to

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

and

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

At $\theta = 0$, taking $\theta \to U_{\alpha\beta}$ and $\theta' \to W_{\beta\gamma}$ gives

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

But, taking $\theta \to U_{\alpha\beta}$ and $\theta' \to b_\beta$ gives

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

So,

$$S_2(\theta) = \frac{1}{2}\sum_{\alpha,\gamma} \frac{J_{\alpha\gamma}}{2M} \sum_\beta U_{\alpha\beta} W_{\beta\gamma} \tag{6}$$

where $J_{\alpha\gamma} := \mathbb{E}_{z\sim p(z)}[z_\gamma] - \mathbb{E}_{z\sim p(z|\alpha)}[z_\gamma]$.

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 $J_{\alpha\gamma}/(2M)$. Computed on fashion-MNIST dataset, components of J vary from -0.1 to 0.075.

We further analyzed S_2 on the best fit $\theta_\star$, trained on training data and evaluated on test data of fashion-MNIST dataset. We found that it is the second term that dominates $S_2(\theta_\star)$. Interestingly, both the terms $J_{\alpha\gamma}$ and $\sum_\beta U_{\alpha\beta} W_{\beta\gamma}$, as rank-2 tensors, have Gaussian distributed elements, centered at zero. But, the multiplied, $J_{\alpha\gamma} \sum_\beta U_{\alpha\beta} W_{\beta\gamma}$, has highly biased elements, most of which are negative. This terms represents the correlation between an output class and a single input dimension.

2.4 Third Order

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

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

To calculate $(\partial^3 S/\partial\theta\partial\theta'\partial\theta'')(0)$, we have to calculate $(\partial^3 f/\partial\theta\partial\theta'\partial\theta'')(z,0)$. Since $\sigma(0)=0$, $\sigma'(0)=1/2$, and $\sigma''(0)=1/2$, we have the non-vanishing terms

$$\begin{aligned}\frac{\partial^3 f_\alpha}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}}(z,0) &= \frac{z_\gamma z_\delta}{2};\\ \frac{\partial^3 f_\alpha}{\partial U_{\alpha\beta}\partial W_{\beta\gamma} b_\beta}(z,0) &= \frac{z_\gamma}{2};\\ \frac{\partial^3 f_\alpha}{\partial U_{\alpha\beta}\partial b_\beta \partial b_\beta}(z,\theta) &= \frac{1}{2}.\end{aligned}$$

Thus, taking $\theta \to U_{\alpha\beta}$, $\theta' \to W_{\beta\gamma}$ and $\theta'' \to W_{\beta\delta}$ gives

$$\begin{aligned}\frac{\partial^3 S}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}}(0) &= \sum_{z,m} p(z,m) \sum_{\alpha'} q_{\alpha'} \frac{\partial^3 f_{\alpha'}}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}} - \sum_{z,m} p(z,m) \frac{\partial^3 f_m}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}}\\ \left\{q_\alpha \equiv \frac{1}{M}, \frac{\partial^3 f}{\partial U \partial W \partial W} = \cdots\right\} &= \sum_{z,m} p(z,m) \sum_{\alpha'} \frac{1}{M} \frac{\delta_{\alpha\alpha'} z_\gamma z_\delta}{2} - \sum_{z,m} p(z,m) \frac{\delta_{m\alpha} z_\gamma z_\delta}{2}\\ \left\{p(\alpha) \equiv \frac{1}{M}\right\} &= \frac{1}{2M} J_{\alpha\gamma\delta}\end{aligned}$$

where $J_{\alpha\gamma\delta} := \mathbb{E}_{z\sim p(z)}[z_\gamma z_\delta] - \mathbb{E}_{z\sim p(z|\alpha)}[z_\gamma z_\delta]$. Following the same process, we find

$$\frac{\partial^3 S}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial b_\beta}(0) = \frac{1}{6M}(\mathbb{E}_{z\sim p(z)}[z_\gamma] - \mathbb{E}_{z\sim p(z|\alpha)}[z_\gamma]) = \frac{1}{2M} J_{\alpha\gamma}$$

and

$$\frac{\partial^3 S}{\partial U_{\alpha\beta}\partial b_\beta \partial b_\beta}(0) = 0.$$

So,

$$S_3(\theta) = \sum_{\alpha,\gamma} \frac{J_{\alpha\gamma}}{12M} \sum_\beta U_{\alpha\beta} W_{\beta\gamma} b_\beta + \sum_{\alpha,\gamma,\delta} \frac{J_{\alpha\gamma\delta}}{12M} \sum_\beta U_{\alpha\beta} W_{\beta\gamma} W_{\beta\delta}.$$

We analyzed S_3 on the best fit $\theta_\star$. We found that it is the last term that dominates $S_3(\theta_\star)$. Interestingly, like the case of S_2, both the terms $J_{\alpha\gamma\delta}$ and $\sum_\beta U_{\alpha\beta} W_{\beta\gamma} W_{\beta\delta}$, as rank-3 tensors, have Gaussian distributed elements, centered at zero. But, the multiplied, $J_{\alpha\gamma\delta} U_{\alpha\beta} W_{\beta\gamma} W_{\beta\delta}$, has highly biased elements, most of which are positive. This terms represents the correlation between an output class and two input dimensions.

Why does the last term dominate S_3? Comparing with other terms, the sub-terms involved in the summation is much more. For example, when U is 10×2048 and W is 2048×1024, the last summation has 2.2×10^{10} sub-terms, other terms have 10^3, 2.1×10^5, 2.1×10^8, and 2.1×10^7 sub-terms, respectively. So, if the scales of U, W, and b are in the same order, then the last term dominates. We can check this idea by making the embedding dimension E small. Indeed, when E is small, domination of the last term vanishes. Notice that the power law between the model size and the optimized loss appears only when E has been large enough. So, we can guess that this domination is the key to the power law.

The problem left is why the scales of U, W, and b are in the same order when $\theta \approx \theta_\star$.

2.5 Higher Orders and Summary

Based on the previous analysis, it is suspected that the main contribution from $S_{n+1}(\theta_\star)$ to $S(\theta_\star)$ is

$$\frac{\sigma^{(n)}(0)}{(n+1)!M} \sum_{\alpha,\gamma_1,\ldots,\gamma_n} J_{\alpha\gamma_1\cdots\gamma_n} \sum_\beta U_{\alpha\beta} W_{\beta\gamma_1} \cdots W_{\beta\gamma_n}$$

where we have defined $J_{\alpha\gamma_1\cdots\gamma_n} := \mathbb{E}_{z\sim p(z)}[z_{\gamma_1}\cdots z_{\gamma_n}] - \mathbb{E}_{z\sim p(z|\alpha)}[z_{\gamma_1}\cdots z_{\gamma_n}]$ as usual. The term $\sum_\beta U_{\alpha\beta} W_{\beta\gamma_1}\cdots W_{\beta\gamma_n}$ characterizes the correlation between an output class α and the input dimensions $\gamma_1,\ldots,\gamma_n$. If this is true, then we have

$$S(\theta) \approx \ln M + \sum_{n=1}^{+\infty} \frac{\sigma^{(n)}(0)}{(n+1)!M} \sum_{\alpha,\gamma_1,\ldots,\gamma_n} J_{\alpha\gamma_1\cdots\gamma_n} \sum_\beta U_{\alpha\beta} W_{\beta\gamma_1}\cdots W_{\beta\gamma_n}$$

for any $\theta \approx \theta_\star$.[2]

3 Data Size and Early Stopping

In fact, we have only finite size of dataset. We cannot get the $p(z,m)$, but empirical distributions $p_T(z,m)$ and $p_E(z,m)$, both of which are summations of delta functions. The p_T for training data and p_V for test (or validation) data. The strategy training is minimizing the action (training loss)

$$S_T(\theta) := -\sum_{z,m} p_T(z,m) \ln q_m(z,\theta)$$

by gradient descent method is optimizing until another action (validation loss)

$$S_V(\theta) := -\sum_{z,m} p_V(z,m) \ln q_m(z,\theta)$$

starts to increase. In this situation, we have $\nabla S_T \cdot \nabla S_V = 0$, where the ∇S_V starts to turn its direction to go against with the ∇S_T. So, the training *early stops* at

$$\nabla S_T(\theta) \cdot \nabla S_V(\theta) = 0, \tag{7}$$

instead of $\nabla S(\theta) = 0$. This difference is especially important when the data size is quite limited.

2. If there is an additional hidden layer, with weights V, thus an educated guess would be proportional to

$$S_{n+2} \propto J_{\alpha\gamma_1\cdots\gamma_n} \sum_{\beta,\beta'} U_{\alpha\beta} V_{\beta\beta'} W_{\beta'\gamma_1}\cdots W_{\beta'\gamma_n}.$$

But it seems that there is degenerancy between U and W. But, there will also be contribution like

$$S_{n+m+2} \propto J_{\alpha\gamma_1\cdots\gamma_n} \sum_{\beta,\beta'} U_{\alpha\beta} V_{\beta\beta'_1}\cdots V_{\beta\beta'_2} W_{\beta'_1\gamma_1}\cdots W_{\beta'_m\gamma_n}.$$