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 $q_m(z, \theta) := \operatorname{softmax}_m(f(z, \theta))$ 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 SiLU activation, that is, $\sigma(x) = x / (1 + \mathrm{e}^{-x})$.

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 $S(\theta_\star) = 0$, where $q_m(z, \theta_\star) = 1$ for each z.

Assume that $p(m) := \sum_z p(z, m) = 1/M$ for all $m = 1, \ldots, M$, 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 $q_\alpha(z, 0) = \operatorname{softmax}_\alpha(f(z, 0)) = 1/M$ for all $\alpha = 1, \ldots, M$. So,

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

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, c, W, and b respectively, we have the non-vanishing terms

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

Setting $\theta=0$, only

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

is left. Thus, we shall take $\theta \to c_\alpha$, that is,

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

So,

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

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

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

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

and

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

At $\theta=0$, taking $\theta\to c_\alpha$ and $\theta'\to c_\beta$ gives

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

Taking $\theta\to U_{\alpha\beta}$ and $\theta'\to W_{\beta\gamma}$ gives

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

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,

$$S_2(\theta)=\frac{1}{2}\sum_{\alpha,\beta}\left(\frac{\delta_{\alpha\beta}}{M}-\frac{1}{M^2}\right)c_\alpha c_\beta + \frac{1}{2}\sum_{\alpha,\gamma}\frac{J_{\alpha\gamma}}{2M}\sum_\beta U_{\alpha\beta}W_{\beta\gamma} \tag{6}$$

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

By numerical computation, we find that the matrix $\delta/M-1/M^2$ is non-positive definite since it has non-positive determinant. 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.

But, numerical computation shows that there is not lower bound for the second term of S_2. This means, the $|U|$ and $|W|$ grows until the next order takes part in. And the matrix $\delta_{\alpha\beta} - 1/M$ has non-positive determinant for $M = 2, 3, \dots$, which means $c = 0$ is also unstable. So, we have to consider the third order.

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 $\sigma''(0) = 1/6$, we have the non-vanishing terms

$$\begin{gathered}
\frac{\partial^3 f_\alpha}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}}(z,0) = \frac{z_\gamma z_\delta}{6}; \\
\frac{\partial^3 f_\alpha}{\partial U_{\alpha\beta}\partial W_{\beta\gamma} b_\beta}(z,0) = \frac{z_\gamma}{6}; \\
\frac{\partial^3 f_\alpha}{\partial U_{\alpha\beta}\partial b_\beta\partial b_\beta}(z,\theta) = \frac{1}{6}.
\end{gathered}$$

Thus, taking $\theta \to c_\alpha$, $\theta' \to c_\beta$ and $\theta'' \to c_\gamma$ gives

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

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

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

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

$$\frac{\partial^3 S}{\partial U_{\alpha\beta}\partial b_\beta \partial c_\gamma}(0) = \frac{\delta_{\alpha\gamma}}{2M} - \frac{1}{2M^2}.$$

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

$$\begin{aligned}
\frac{\partial^3 S}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}}(0) &= \sum_{z,m} p(z,m) \sum_{\alpha'} q_{\alpha'} \frac{\partial^3 f_{\alpha'}}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}} - \sum_{z,m} p(z,m) \frac{\partial^3 f_m}{\partial U_{\alpha\beta}\partial W_{\beta\gamma}\partial W_{\beta\delta}} \\
\left\{ q_\alpha \equiv \frac{1}{M'} \frac{\partial^3 f}{\partial U \partial W \partial W} = \cdots \right\} &= \sum_{z,m} p(z,m) \sum_{\alpha'} \frac{1}{M} \frac{\delta_{\alpha\alpha'} z_\gamma z_\delta}{6} - \sum_{z,m} p(z,m) \frac{\delta_{m\alpha} z_\gamma z_\delta}{6} \\
\left\{ p(\alpha) \equiv \frac{1}{M} \right\} &= \frac{1}{6M} J_{\alpha\gamma\delta}
\end{aligned}$$

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

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

and

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

So,

$$\begin{aligned}
S_3(\theta) = &\sum_{\alpha,\beta,\gamma} \left(\frac{\delta_{\alpha\beta\gamma}}{6M} - \frac{\delta_{\alpha\beta} + \delta_{\alpha\gamma} + \delta_{\beta\gamma}}{6M^2} + \frac{1}{3M^3} \right) c_\alpha c_\beta c_\gamma \\
&+ \sum_{\alpha,\beta,\gamma} \left(\frac{\delta_{\alpha\gamma}}{12M} - \frac{1}{12M^2} \right) U_{\alpha\beta} b_\beta c_\gamma \\
&+ \sum_{\alpha,\beta,\gamma,\delta} \left(\frac{Z_\gamma \delta_{\alpha\delta}}{12M} - \frac{Z_\gamma}{12M^2} \right) U_{\alpha\beta} W_{\beta\gamma} c_\delta \\
&+ \sum_{\alpha,\beta,\gamma} \frac{J_{\alpha\gamma}}{36M} U_{\alpha\beta} W_{\beta\gamma} b_\beta \\
&+ \sum_{\alpha,\beta,\gamma,\delta} \frac{J_{\alpha\gamma\delta}}{36M} U_{\alpha\beta} W_{\beta\gamma} W_{\beta\delta}.
\end{aligned}$$

Numerical computation again shows that, up to the third order, the action still has no lower bound.

We further analyzed S_3 on the best fit $\theta_\star$. We found that it is the last term that dominates $S_2(\theta_\star)$. 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.

2.5 Higher Orders

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

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

which characterizes the correlation between an output class α and the input dimensions $\gamma_1, \ldots, \gamma_n$, and $J_{\alpha\gamma_1\cdots\gamma_n} := \mathbb{E}_{z\sim p(z)}[z_{\gamma_1}\cdots z_{\gamma_n}] - \mathbb{E}_{z\sim p(z|\alpha)}[z_{\gamma_1}\cdots z_{\gamma_n}]$ as usual.