A Proof of Proposition 1

We provide here a detailed proof of Proposition 1.

A.1 Forward Propagation

The optimization problem is

$$\mathsf{csoftmax}(\boldsymbol{z}, \boldsymbol{u}) = \mathsf{argmin} \quad -H(\boldsymbol{\alpha}) - \boldsymbol{z}^\top \boldsymbol{\alpha} \quad \text{s.t.} \quad \begin{cases} \mathbf{1}^\top \boldsymbol{\alpha} = 1 \\ \mathbf{0} \leq \boldsymbol{\alpha} \leq \boldsymbol{u}. \end{cases}$$

The Lagrangian function is:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \lambda, \boldsymbol{\mu}, \boldsymbol{\nu}) &= -H(\boldsymbol{\alpha}) - \boldsymbol{z}^\top \boldsymbol{\alpha} + \lambda(\mathbf{1}^\top \boldsymbol{\alpha} - 1) \\ &\quad -\boldsymbol{\mu}^\top \boldsymbol{\alpha} + \boldsymbol{\nu}^\top (\boldsymbol{\alpha} - \boldsymbol{u}). \end{aligned} \tag{14}$$

To obtain the solution, we invoke the Karush-Kuhn-Tucker conditions. From the stationarity condition, we have $\mathbf{0} = \log(\boldsymbol{\alpha}) + \mathbf{1} - \boldsymbol{z} + \lambda \mathbf{1} - \boldsymbol{\mu} + \boldsymbol{\nu}$, which due to the primal feasibility condition implies that the solution is of the form:

$$\boldsymbol{\alpha} = \exp(\boldsymbol{z} + \boldsymbol{\mu} - \boldsymbol{\nu})/Z, \tag{15}$$

where Z is a normalization constant. From the complementarity slackness condition, we have that $0 < \alpha_i < u_i$ implies that $\mu_i = \nu_i = 0$ and therefore $\alpha_i = \exp(z_i)/Z$. On the other hand, $\nu_i > 0$ implies $\alpha_i = u_i$. Hence the solution can be written as $\alpha_i = \min\{\exp(z_i)/Z, u_i\}$, where Z is determined such that the distribution normalizes:

$$Z = \frac{\sum_{i \in \mathcal{A}} \exp(z_i)}{1 - \sum_{i \notin \mathcal{A}} u_i}, \tag{16}$$

with $\mathcal{A} = \{i \in [L] \mid \alpha_i < u_i\}$.

A.2 Gradient Backpropagation

We now turn to the problem of backpropagating the gradients through the constrained softmax transformation. For that, we need to compute its Jacobian matrix, i.e., the derivatives $\frac{\partial \alpha_i}{\partial z_j}$ and $\frac{\partial \alpha_i}{\partial u_j}$ for $i, j \in [L]$. Let us first express $\boldsymbol{\alpha}$ as

$$\alpha_i = \begin{cases} \frac{\exp(z_i)(1-s)}{\sum_{j \in \mathcal{A}} \exp(z_j)}, & i \in \mathcal{A} \\ u_i, & i \notin \mathcal{A}, \end{cases} \tag{17}$$

where $s = \sum_{j \notin \mathcal{A}} u_j$. Note that we have $\partial s / \partial z_j = 0, \ \forall j$, and $\partial s / \partial u_j = \mathbb{1}(j \notin \mathcal{A})$. To compute the entries of the Jacobian matrix, we need to consider several cases.

Case 1: $\boxed{i \in \mathcal{A}.}$ In this case, the evaluation of Eq. 17 goes through the first branch. Let us first compute the derivative with respect to u_j. Two things can happen: if $j \in \mathcal{A}$, then s does not depend on u_j, hence $\frac{\partial \alpha_i}{\partial u_j} = 0$. Else, if $j \notin \mathcal{A}$, we have

$$\frac{\partial \alpha_i}{\partial u_j} = \frac{-\exp(z_i) \frac{\partial s}{\partial u_j}}{\sum_{k \in \mathcal{A}} \exp(z_k)} = -\alpha_i/(1-s).$$

Now let us compute the derivative with respect to z_j. Three things can happen: if $j \in \mathcal{A}$ and $i \neq j$, we have

$$\begin{aligned} \frac{\partial \alpha_i}{\partial z_j} &= \frac{-\exp(z_i)\exp(z_j)(1-s)}{\left(\sum_{k \in \mathcal{A}} \exp(z_k)\right)^2} \\ &= -\alpha_i \alpha_j/(1-s). \end{aligned} \tag{18}$$

If $j \in \mathcal{A}$ and $i = j$, we have

$$\begin{aligned}\frac{\partial \alpha_i}{\partial z_i} &= (1-s) \times \\ &\quad \frac{\mathsf{exp}(z_i) \sum_{k \in \mathcal{A}} \mathsf{exp}(z_k) - \mathsf{exp}(z_i)^2}{\left(\sum_{k \in \mathcal{A}} \mathsf{exp}(z_k)\right)^2} \\ &= \alpha_i - \alpha_i^2/(1-s). \end{aligned} \tag{19}$$

Finally, if $j \notin \mathcal{A}$, we have $\frac{\partial \alpha_i}{\partial z_j} = 0$.

Case 2: $\boxed{i \notin \mathcal{A}.}$ In this case, the evaluation of Eq. 17 goes through the second branch, which means that $\frac{\partial \alpha_i}{\partial z_j} = 0$, always. Let us now compute the derivative with respect to u_j. This derivative is always zero unless $i = j$, in which case $\frac{\partial \alpha_i}{\partial u_j} = 1$.

To sum up, we have:

$$\frac{\partial \alpha_i}{\partial z_j} = \begin{cases} \mathbb{1}(i=j)\alpha_i - \frac{\alpha_i \alpha_j}{1-s}, & \text{if } i, j \in \mathcal{A} \\ 0, & \text{otherwise,} \end{cases} \tag{20}$$

and

$$\frac{\partial \alpha_i}{\partial u_j} = \begin{cases} -\frac{\alpha_i}{1-s}, & \text{if } i \in \mathcal{A}, j \notin \mathcal{A} \\ 1, & \text{if } i, j \notin \mathcal{A}, i = j \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Therefore, we obtain:

$$\begin{aligned} \mathrm{d}z_j &= \sum_i \frac{\partial \alpha_i}{\partial z_j} \mathrm{d}\alpha_i \\ &= \mathbb{1}(j \in \mathcal{A}) \left(\alpha_j \mathrm{d}\alpha_j - \frac{\alpha_j \sum_{i \in \mathcal{A}} \alpha_i \mathrm{d}\alpha_i}{1-s} \right) \\ &= \mathbb{1}(j \in \mathcal{A}) \alpha_j (\mathrm{d}\alpha_j - m), \end{aligned} \tag{22}$$

and

$$\begin{aligned} \mathrm{d}u_j &= \sum_i \frac{\partial \alpha_i}{\partial u_j} \mathrm{d}\alpha_i \\ &= \mathbb{1}(j \notin \mathcal{A}) \left(\mathrm{d}\alpha_j - \frac{\sum_{i \in \mathcal{A}} \alpha_i \mathrm{d}\alpha_i}{1-s} \right) \\ &= \mathbb{1}(j \notin \mathcal{A}) (\mathrm{d}\alpha_j - m), \end{aligned} \tag{23}$$

where $m = \frac{\sum_{i \in \mathcal{A}} \alpha_i \mathrm{d}\alpha_i}{1-s}$.