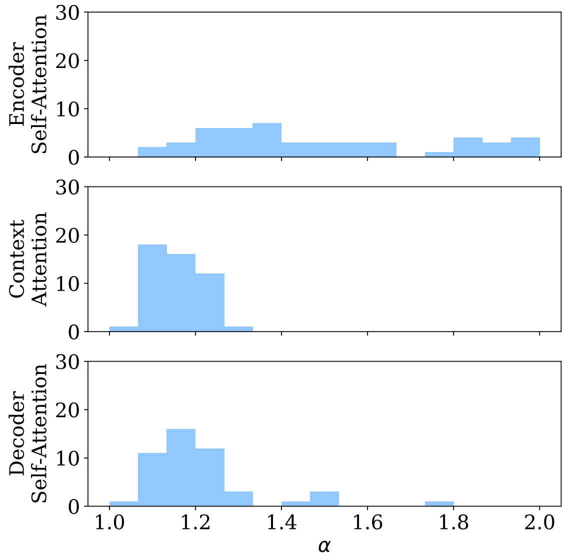
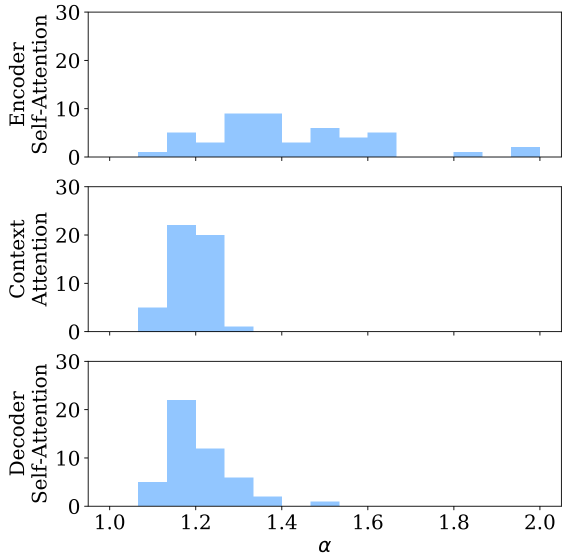
Supplementary Material

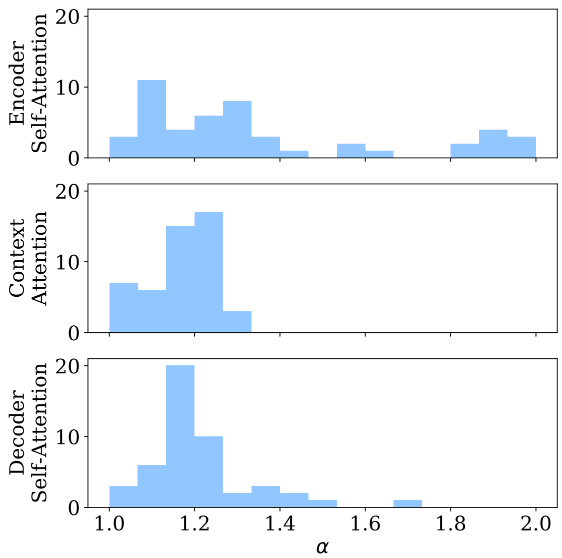
A High-Level Statistics Analysis of Other Language Pairs



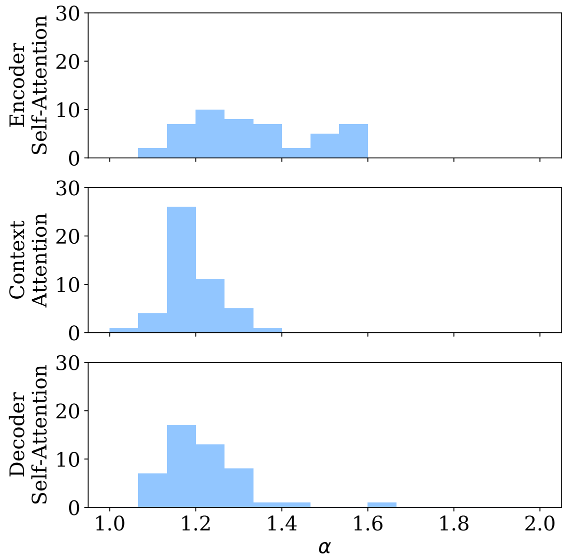
(a) WMT 2016 RO→EN.



(b) KFTT JA→EN.

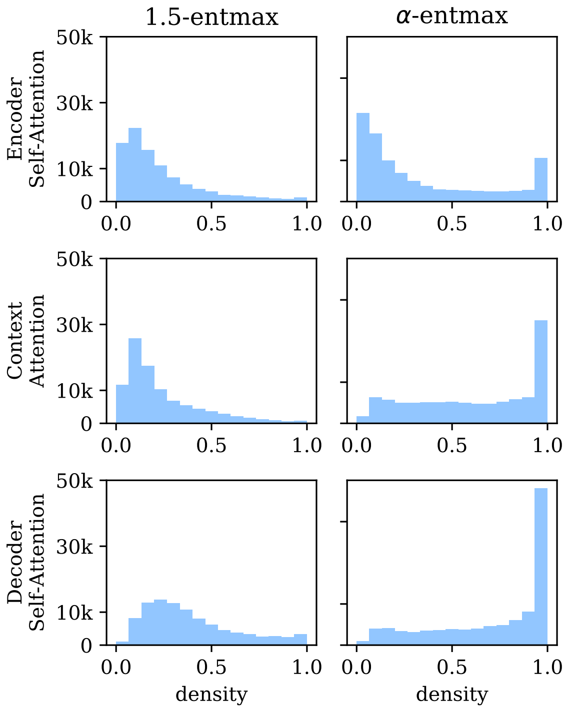


(c) WMT 2014 EN→DE.

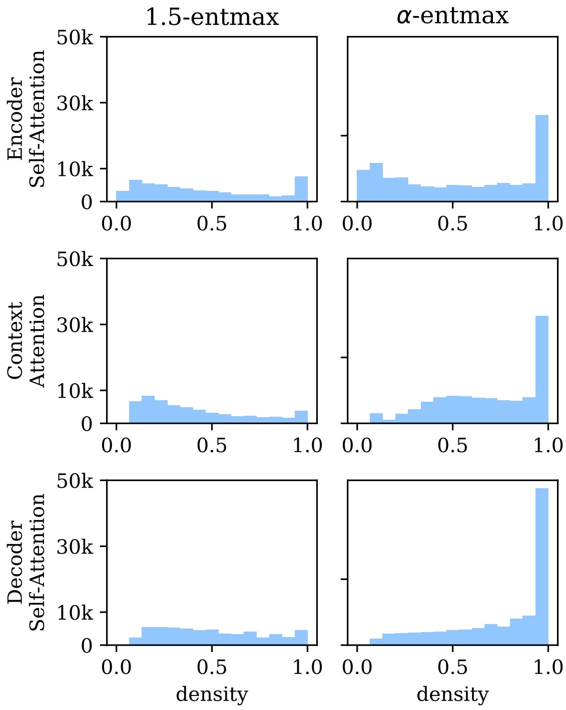


(d) IWSLT 2017 DE→EN.

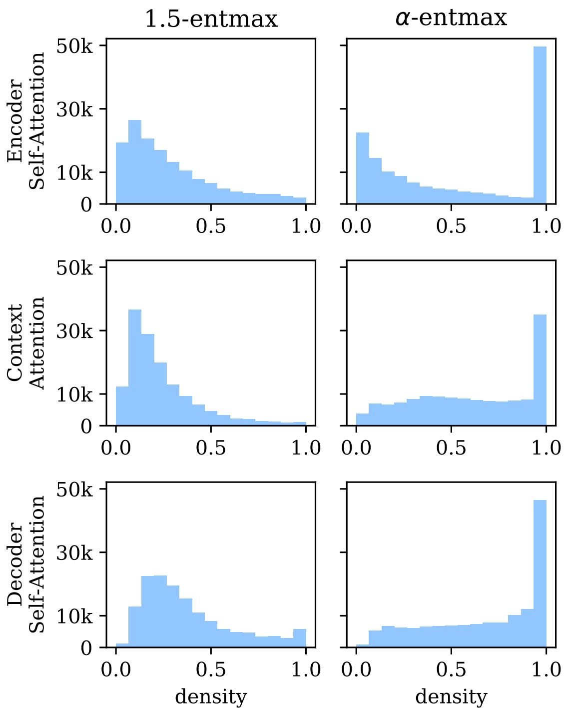
Figure 11: Histograms of α values.



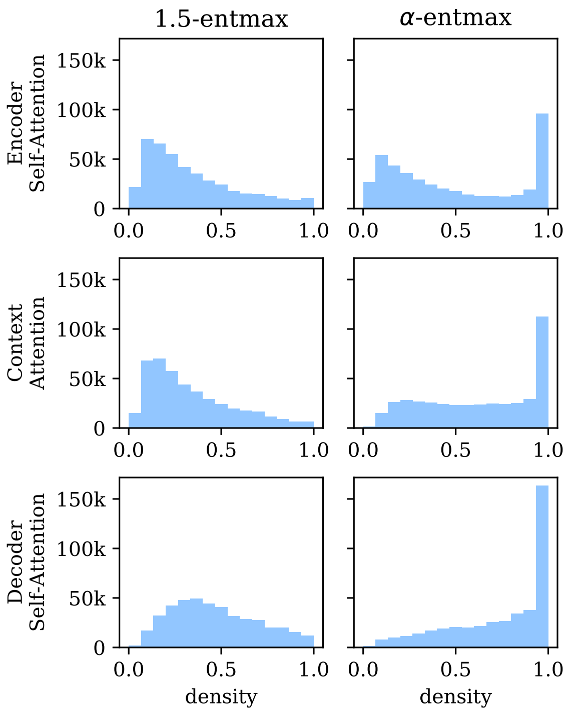
(a) WMT 2016 RO→EN.



(b) KFTT JA→EN.

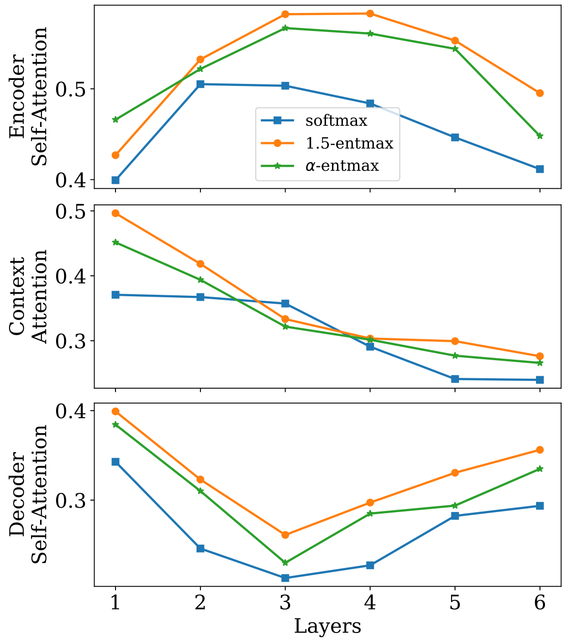


(c) WMT 2014 EN→DE.

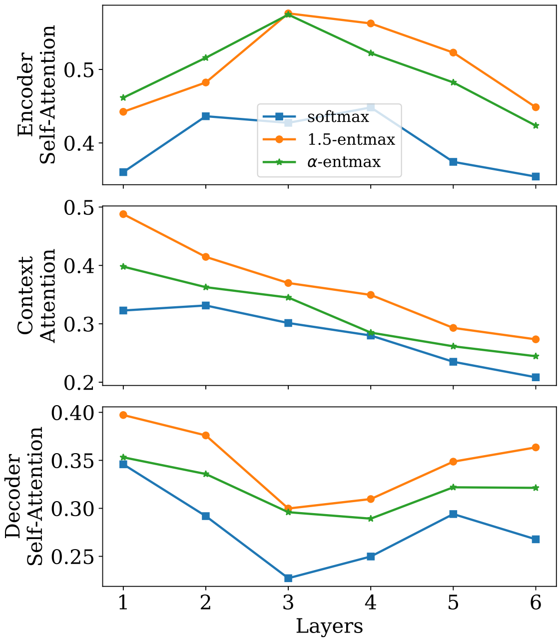


(d) IWSLT 2017 DE→EN.

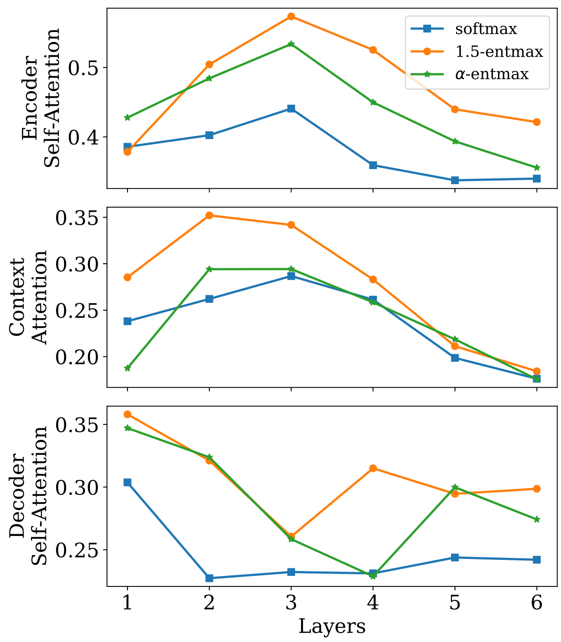
Figure 12: Histograms of head densities.



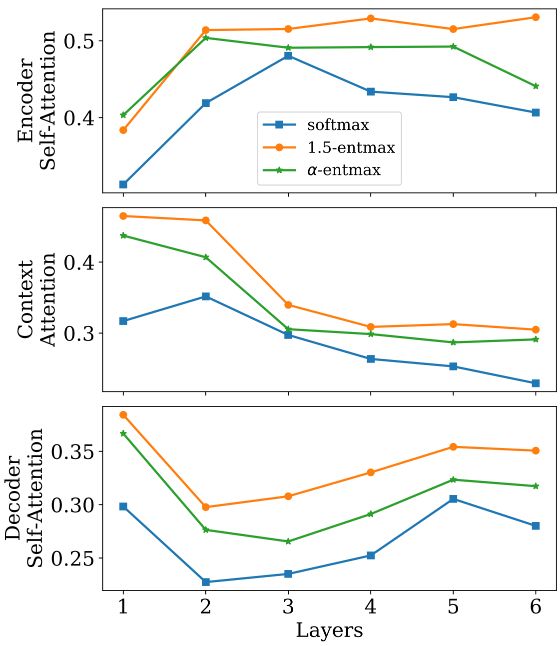
(a) WMT 2016 RO→EN.



(b) KFTT JA→EN.

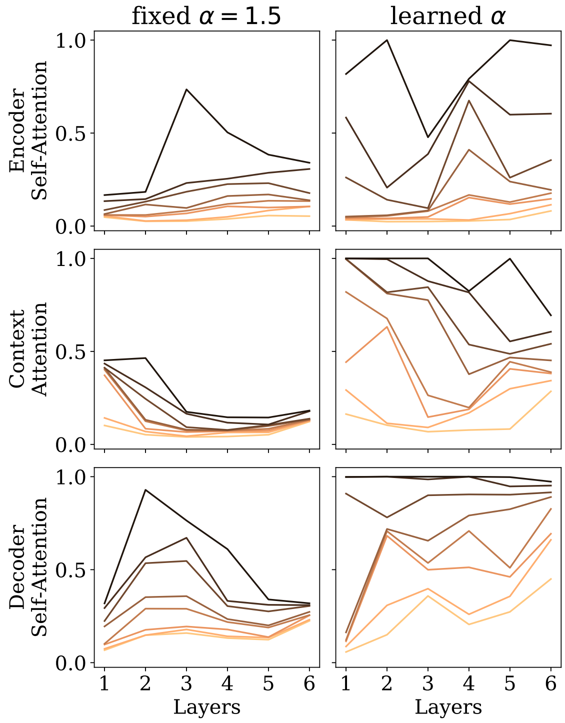


(c) WMT 2014 EN→DE.

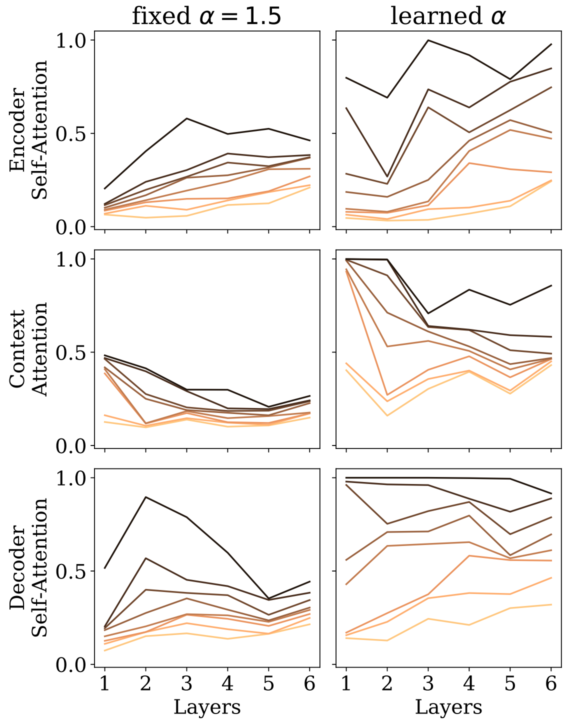


(d) IWSLT 2017 DE→EN.

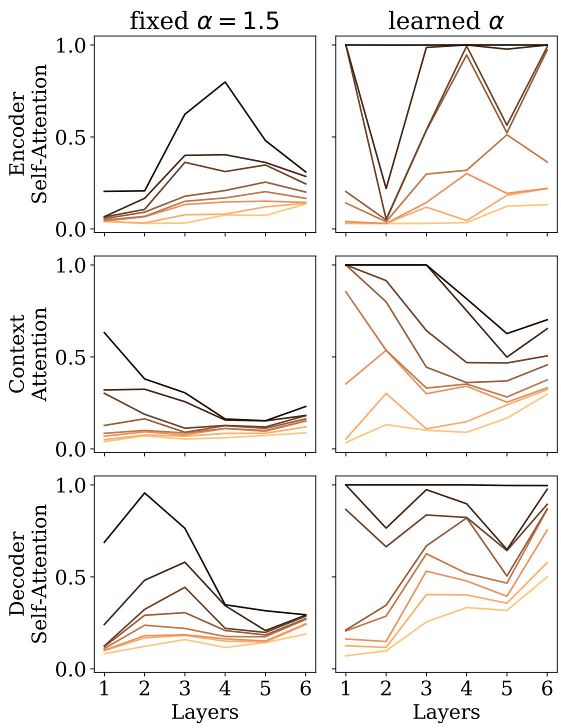
Figure 13: Jensen-Shannon divergence over layers.



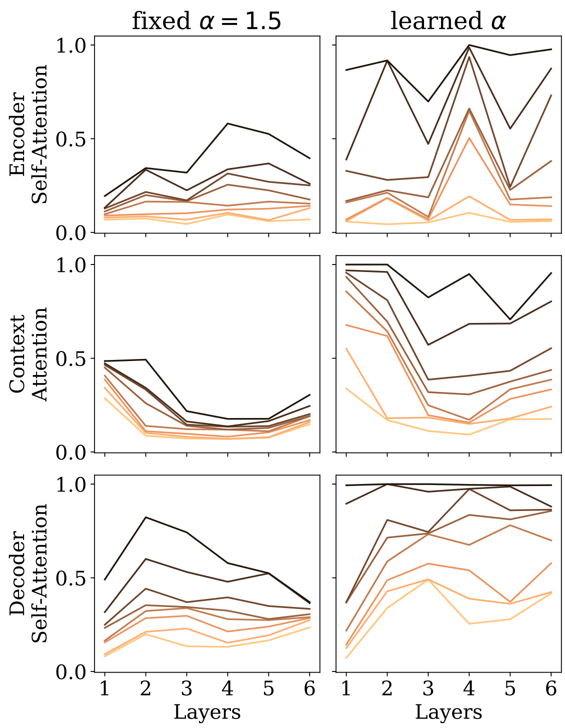
(a) WMT 2016 RO→EN.



(b) KFTT JA→EN.



(c) WMT 2014 EN→DE.



(d) IWSLT 2017 DE→EN.

Figure 14: Head densities over layers.

B Background

B.1 Regularized Fenchel-Young prediction functions

Definition 1 (Blondel et al. 2019)**.** *Let* $\Omega\colon \triangle^d \to \mathbb{R} \cup \{\infty\}$ *be a strictly convex regularization function. We define the prediction function* $\boldsymbol{\pi}_\Omega$ *as*

$$\boldsymbol{\pi}_\Omega(\boldsymbol{z}) = \underset{\boldsymbol{p}\in\triangle^d}{\mathsf{argmax}} \left(\boldsymbol{p}^\top \boldsymbol{z} - \Omega(\boldsymbol{p})\right) \tag{12}$$

B.2 Characterizing the α-entmax mapping

Lemma 1 (Peters et al. 2019)**.** *For any* $\boldsymbol{z}$*, there exists a unique* $\tau^\star$ *such that*

$$\alpha\text{-}\mathsf{entmax}(\boldsymbol{z}) = [(\alpha-1)\boldsymbol{z} - \tau^\star \mathbf{1}]_+^{1/\alpha-1}. \tag{13}$$

Proof: From the definition of α-entmax,

$$\alpha\text{-}\mathsf{entmax}(\boldsymbol{z}) := \underset{\boldsymbol{p}\in\triangle^d}{\mathsf{argmax}}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathsf{H}_\alpha^\mathsf{T}(\boldsymbol{p}), \tag{14}$$

we may easily identify it with a regularized prediction function (Def. 1):

$$\alpha\text{-}\mathsf{entmax}(\boldsymbol{z}) \equiv \boldsymbol{\pi}_{-\mathsf{H}_\alpha^\mathsf{T}}(\boldsymbol{z}).$$

We first note that for all $\boldsymbol{p} \in \triangle^d$,

$$-(\alpha-1)\mathsf{H}_\alpha^\mathsf{T}(\boldsymbol{p}) = \frac{1}{\alpha}\sum_{i=1}^d p_i^\alpha + \text{const}. \tag{15}$$

From the constant invariance and scaling properties of $\boldsymbol{\pi}_\Omega$ (Blondel et al., 2019, Proposition 1, items 4–5),

$$\boldsymbol{\pi}_{-\mathsf{H}_\alpha^\mathsf{T}}(\boldsymbol{z}) = \boldsymbol{\pi}_\Omega((\alpha-1)\boldsymbol{z}), \quad \text{with} \quad \Omega(\boldsymbol{p}) = \sum_{j=1}^d g(p_j), \quad g(t) = \frac{t^\alpha}{\alpha}.$$

Using (Blondel et al., 2019, Proposition 5), noting that $g'(t) = t^{\alpha-1}$ and $(g')^{-1}(u) = u^{1/\alpha-1}$, yields

$$\boldsymbol{\pi}_\Omega(\boldsymbol{z}) = [\boldsymbol{z} - \tau^\star\mathbf{1}]_+^{1/\alpha-1}, \quad \text{and therefore} \quad \alpha\text{-}\mathsf{entmax}(\boldsymbol{z}) = [(\alpha-1)\boldsymbol{z} - \tau^\star\mathbf{1}]_+^{1/\alpha-1}. \tag{16}$$

Since $\mathsf{H}_\alpha^\mathsf{T}$ is strictly convex on the simplex, α-entmax has a unique solution $\boldsymbol{p}^\star$. Equation 16 implicitly defines a one-to-one mapping between $\boldsymbol{p}^\star$ and $\tau^\star$ as long as $\boldsymbol{p}^\star \in \triangle$, therefore $\tau^\star$ is also unique. ■

B.3 Connections to softmax and sparsemax

The Euclidean projection onto the simplex, sometimes referred to, in the context of neural attention, as sparsemax (Martins and Astudillo, 2016), is defined as

$$\mathsf{sparsemax}(\boldsymbol{z}) := \underset{\boldsymbol{p}\in\triangle}{\mathsf{argmin}}\, \|\boldsymbol{p} - \boldsymbol{z}\|_2^2. \tag{17}$$

The solution can be characterized through the unique threshold τ such that $\sum_i \mathsf{sparsemax}(\boldsymbol{z})_i = 1$ and (Held et al., 1974)

$$\mathsf{sparsemax}(\boldsymbol{z}) = [\boldsymbol{z} - \tau\mathbf{1}]_+\,. \tag{18}$$

Thus, each coordinate of the sparsemax solution is a piecewise-linear function. Visibly, this expression is recovered when setting $\alpha = 2$ in the α-entmax expression (Equation 21); for other values of α, the exponent induces curvature.

On the other hand, the well-known softmax is usually defined through the expression

$$\mathsf{softmax}(\boldsymbol{z})_i := \frac{\exp(z_i)}{\sum_j \exp(z_j)}, \tag{19}$$

which can be shown to be the unique solution of the optimization problem

$$\mathsf{softmax}(\boldsymbol{z})_i = \underset{\boldsymbol{p} \in \triangle}{\mathsf{argmax}}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathsf{H}^{\mathsf{S}}(\boldsymbol{p}), \tag{20}$$

where $\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}) := -\sum_i p_i \log p_i$ is the Shannon entropy. Indeed, setting the gradient to 0 yields the condition $\log p_i = z_j - \nu_i - \tau - 1$, where τ and $\nu > 0$ are Lagrange multipliers for the simplex constraints $\sum_i p_i = 1$ and $p_i \geq 0$, respectively. Since the *l.h.s.* is only finite for $p_i > 0$, we must have $\nu_i = 0$ for all i, by complementary slackness. Thus, the solution must have the form $p_i = {\exp(z_i)}/{Z}$, yielding Equation 19.

C Jacobian of α-entmax *w.r.t.* the shape parameter α: Proof of Proposition 1

Recall that the entmax transformation is defined as:

$$\alpha\text{-}\mathsf{entmax}(\boldsymbol{z}) := \underset{\boldsymbol{p} \in \triangle^d}{\mathsf{argmax}}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathsf{H}^{\mathsf{T}}_{\alpha}(\boldsymbol{p}), \tag{21}$$

where $\alpha \geq 1$ and $\mathsf{H}^{\mathsf{T}}_{\alpha}$ is the Tsallis entropy,

$$\mathsf{H}^{\mathsf{T}}_{\alpha}(\boldsymbol{p}) := \begin{cases} \frac{1}{\alpha(\alpha-1)} \sum_j \left(p_j - p_j^{\alpha}\right), & \alpha \neq 1, \\ \mathsf{H}^{\mathsf{S}}(\boldsymbol{p}), & \alpha = 1, \end{cases} \tag{22}$$

and $\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}) := -\sum_j p_j \log p_j$ is the Shannon entropy.

In this section, we derive the Jacobian of entmax with respect to the scalar parameter α.

C.1 General case of $\alpha > 1$

From the KKT conditions associated with the optimization problem in Eq. 21, we have that the solution $\boldsymbol{p}^\star$ has the following form, coordinate-wise:

$$p_i^\star = [(\alpha - 1)(z_i - \tau^\star)]_+^{1/(\alpha-1)}, \tag{23}$$

where $\tau^\star$ is a scalar Lagrange multiplier that ensures that $\boldsymbol{p}^\star$ normalizes to 1, *i.e.*, it is defined implicitly by the condition:

$$\sum_i [(\alpha - 1)(z_i - \tau^\star)]_+^{1/(\alpha-1)} = 1. \tag{24}$$

For general values of α, Eq. 24 lacks a closed form solution. This makes the computation of the Jacobian

$$\frac{\partial\, \alpha\text{-}\mathsf{entmax}(\boldsymbol{z})}{\partial \alpha} \tag{25}$$

non-trivial. Fortunately, we can use the technique of implicit differentiation to obtain this Jacobian.

The Jacobian exists almost everywhere, and the expressions we derive expressions yield a generalized Jacobian (Clarke, 1990) at any non-differentiable points that may occur for certain (α, $\boldsymbol{z}$) pairs. We begin by noting that $\frac{\partial p_i^\star}{\partial \alpha} = 0$ if $p_i^\star = 0$, because increasing α keeps sparse coordinates sparse.[4] Therefore we need to worry only about coordinates that are in the support of $\boldsymbol{p}^\star$. We will assume hereafter that the i^{th} coordinate of $\boldsymbol{p}^\star$ is non-zero. We have:

$$\begin{aligned}
\frac{\partial p_i^\star}{\partial \alpha} &= \frac{\partial}{\partial \alpha}[(\alpha-1)(z_i-\tau^\star)]^{\frac{1}{\alpha-1}} \\
&= \frac{\partial}{\partial \alpha}\exp\left[\frac{1}{\alpha-1}\log[(\alpha-1)(z_i-\tau^\star)]\right] \\
&= p_i^\star \frac{\partial}{\partial \alpha}\left[\frac{1}{\alpha-1}\log[(\alpha-1)(z_i-\tau^\star)]\right] \\
&= \frac{p_i^\star}{(\alpha-1)^2}\left[\frac{\frac{\partial}{\partial \alpha}[(\alpha-1)(z_i-\tau^\star)]}{z_i-\tau^\star} - \log[(\alpha-1)(z_i-\tau^\star)]\right] \\
&= \frac{p_i^\star}{(\alpha-1)^2}\left[\frac{z_i-\tau^\star-(\alpha-1)\frac{\partial \tau^\star}{\partial \alpha}}{z_i-\tau^\star} - \log[(\alpha-1)(z_i-\tau^\star)]\right] \\
&= \frac{p_i^\star}{(\alpha-1)^2}\left[1 - \frac{\alpha-1}{z_i-\tau^\star}\frac{\partial \tau^\star}{\partial \alpha} - \log[(\alpha-1)(z_i-\tau^\star)]\right]. \qquad (26)
\end{aligned}$$

We can see that this Jacobian depends on $\frac{\partial \tau^\star}{\partial \alpha}$, which we now compute using implicit differentiation.

Let $\mathcal{S} = \{i : p_i^\star > 0\}$). By differentiating both sides of Eq. 24, re-using some of the steps in Eq. 26, and recalling Eq. 23, we get

$$\begin{aligned}
0 &= \sum_{i\in\mathcal{S}} \frac{\partial}{\partial \alpha}[(\alpha-1)(z_i-\tau^\star)]^{1/(\alpha-1)} \\
&= \sum_{i\in\mathcal{S}} \frac{p_i^\star}{(\alpha-1)^2}\left[1 - \frac{\alpha-1}{z_i-\tau^\star}\frac{\partial \tau^\star}{\partial \alpha} - \log[(\alpha-1)(z_i-\tau^\star)]\right] \\
&= \frac{1}{(\alpha-1)^2} - \frac{\partial \tau^\star}{\partial \alpha}\sum_{i\in\mathcal{S}}\frac{p_i^\star}{(\alpha-1)(z_i-\tau^\star)} - \sum_{i\in\mathcal{S}}\frac{p_i^\star}{(\alpha-1)^2}\log[(\alpha-1)(z_i-\tau^\star)] \\
&= \frac{1}{(\alpha-1)^2} - \frac{\partial \tau^\star}{\partial \alpha}\sum_i (p_i^\star)^{2-\alpha} - \sum_i \frac{p_i^\star}{\alpha-1}\log p_i^\star \\
&= \frac{1}{(\alpha-1)^2} - \frac{\partial \tau^\star}{\partial \alpha}\sum_i (p_i^\star)^{2-\alpha} + \frac{\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^*)}{\alpha-1}, \qquad (27)
\end{aligned}$$

from which we obtain:

$$\frac{\partial \tau^\star}{\partial \alpha} = \frac{\frac{1}{(\alpha-1)^2} + \frac{\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)}{\alpha-1}}{\sum_i (p_i^\star)^{2-\alpha}}. \qquad (28)$$

Finally, plugging Eq. 28 into Eq. 26, we get:

$$\begin{aligned}
\frac{\partial p_i^\star}{\partial \alpha} &= \frac{p_i^\star}{(\alpha-1)^2}\left[1 - \frac{1}{(p_i^\star)^{\alpha-1}}\frac{\partial \tau^\star}{\partial \alpha} - (\alpha-1)\log p_i^\star\right] \\
&= \frac{p_i^\star}{(\alpha-1)^2}\left[1 - \frac{1}{(p_i^\star)^{\alpha-1}}\frac{\frac{1}{(\alpha-1)^2} + \frac{\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)}{\alpha-1}}{\sum_i (p_i^\star)^{2-\alpha}} - (\alpha-1)\log p_i^\star\right] \\
&= \frac{p_i^\star - \tilde{p}_i(\alpha)}{(\alpha-1)^2} - \frac{p_i^\star \log p_i^\star + \tilde{p}_i(\alpha)\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)}{\alpha-1}, \qquad (29)
\end{aligned}$$

[4] This follows from the margin property of $\mathsf{H}^{\mathsf{T}}_\alpha$ (Blondel et al., 2019).

where we denote by

$$\tilde{p}_i(\alpha) = \frac{(p_i^\star)^{2-\alpha}}{\sum_j (p_j^\star)^{2-\alpha}}. \tag{30}$$

The distribution $\tilde{\boldsymbol{p}}(\alpha)$ can be interpreted as a "skewed" distribution obtained from $\boldsymbol{p}^\star$, which appears in the Jacobian of α-$\mathsf{entmax}(\boldsymbol{z})$ *w.r.t.* $\boldsymbol{z}$ as well (Peters et al., 2019).

C.2 Solving the indetermination for $\alpha = 1$

We can write Eq. 29 as

$$\frac{\partial p_i^\star}{\partial \alpha} = \frac{p_i^\star - \tilde{p}_i(\alpha) - (\alpha-1)(p_i^\star \log p_i^\star + \tilde{p}_i(\alpha)\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star))}{(\alpha-1)^2}. \tag{31}$$

When $\alpha \to 1^+$, we have $\tilde{\boldsymbol{p}}(\alpha) \to \boldsymbol{p}^\star$, which leads to a $\frac{0}{0}$ indetermination.

To solve this indetermination, we will need to apply L'Hôpital's rule twice. Let us first compute the derivative of $\tilde{p}_i(\alpha)$ with respect to α. We have

$$\frac{\partial}{\partial \alpha}(p_i^\star)^{2-\alpha} = -(p_i^\star)^{2-\alpha}\log p_i^\star, \tag{32}$$

therefore

$$\begin{aligned}
\frac{\partial}{\partial \alpha}\tilde{p}_i(\alpha) &= \frac{\partial}{\partial \alpha}\frac{(p_i^\star)^{2-\alpha}}{\sum_j (p_j^\star)^{2-\alpha}} \\
&= \frac{-(p_i^\star)^{2-\alpha}\log p_i^\star \sum_j (p_j^\star)^{2-\alpha} + (p_i^\star)^{2-\alpha}\sum_j (p_j^\star)^{2-\alpha}\log p_j^\star}{\left(\sum_j (p_j^\star)^{2-\alpha}\right)^2} \\
&= -\tilde{p}_i(\alpha)\log p_i^\star + \tilde{p}_i(\alpha)\sum_j \tilde{p}_j(\alpha)\log p_j^\star.
\end{aligned} \tag{33}$$

Differentiating the numerator and denominator in Eq. 31, we get:

$$\begin{aligned}
\frac{\partial p_i^\star}{\partial \alpha} &= \lim_{\alpha\to 1^+}\frac{(1+(\alpha-1)\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star))\tilde{p}_i(\alpha)(\log p_i^\star - \sum_j \tilde{p}_j(\alpha)\log p_j^\star) - p_i^\star \log p_i^\star - \tilde{p}_i(\alpha)\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)}{2(\alpha-1)} \\
&= A + B,
\end{aligned} \tag{34}$$

with

$$\begin{aligned}
A &= \lim_{\alpha\to 1^+}\frac{\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)\tilde{p}_i(\alpha)(\log p_i^\star - \sum_j \tilde{p}_j(\alpha)\log p_j^\star)\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)}{2} \\
&= \frac{\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)p_i^\star \log p_i^\star + p_i^\star(\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star))^2}{2},
\end{aligned} \tag{35}$$

and

$$B = \lim_{\alpha\to 1^+}\frac{\tilde{p}_i(\alpha)(\log p_i^\star - \sum_j \tilde{p}_j(\alpha)\log p_j^\star) - p_i^\star \log p_i^\star - \tilde{p}_i(\alpha)\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)}{2(\alpha-1)}. \tag{36}$$

When $\alpha \to 1^+$, B becomes again a $\frac{0}{0}$ indetermination, which we can solve by applying again L'Hôpital's rule. Differentiating the numerator and denominator in Eq. 36:

$$\begin{aligned}
B &= \frac{1}{2}\lim_{\alpha\to 1^+}\Bigg\{\tilde{p}_i(\alpha)\log p_i^\star\left(\sum_j \tilde{p}_j(\alpha)\log p_j^\star - \log p_i^\star\right)\\
&\quad -\tilde{p}_i(\alpha)\left(\sum_j \tilde{p}_j(\alpha)\log p_j^\star - \log p_i^\star\right)\left(\sum_j \tilde{p}_j(\alpha)\log p_j^\star + \mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)\right)\\
&\quad -\tilde{p}_i(\alpha)\sum_j \tilde{p}_j(\alpha)\log p_j^\star\left(\sum_k \tilde{p}_k(\alpha)\log p_k^\star - \log p_j^\star\right)\Bigg\}\\
&= \frac{-p_i^\star\log p_i^\star(\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)+\log p_i^\star) + p_i^\star\sum_j p_j^\star\log p_j^\star(\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)+\log p_j^\star)}{2}\\
&= \frac{-\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)p_i^\star\log p_i^\star - p_i^\star(\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star))^2 - p_i^\star\log^2 p_i^\star + p_i^\star\sum_j p_j^\star\log^2 p_j^\star}{2}. \qquad (37)
\end{aligned}$$

Finally, summing Eq. 35 and Eq. 37, we get

$$\left.\frac{\partial p_i^\star}{\partial\alpha}\right|_{\alpha=1} = \frac{-p_i^\star\log^2 p_i^\star + p_i^\star\sum_j p_j^\star\log^2 p_j^\star}{2}. \qquad (38)$$

C.3 Summary

To sum up, we have the following expression for the Jacobian of α-entmax with respect to α:

$$\frac{\partial p_i^\star}{\partial\alpha} = \begin{cases} \frac{p_i^\star-\tilde{p}_i(\alpha)}{(\alpha-1)^2} - \frac{p_i^\star\log p_i^\star + \tilde{p}_i(\alpha)\mathsf{H}^{\mathsf{S}}(\boldsymbol{p}^\star)}{\alpha-1}, & \text{for } \alpha>1\\ \frac{-p_i^\star\log^2 p_i^\star + p_i^\star\sum_j p_j^\star\log^2 p_j^\star}{2}, & \text{for } \alpha=1. \end{cases} \qquad (39)$$