Supplementary Material

A  High-Level Statistics Analysis of Other Language Pairs

Figure 11: Histograms of $\alpha$ values.
Figure 12: Histograms of head densities.
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.

(a) WMT 2016 RO→EN.
(b) KFTT JA→EN.
(c) WMT 2014 EN→DE.
(d) IWSLT 2017 DE→EN.
B Background

B.1 Regularized Fenchel-Y oung prediction functions

Definition 1 (Blondel et al. 2019). Let \( \Omega : \Delta^d \to \mathbb{R} \cup \{\infty\} \) be a strictly convex regularization function. We define the prediction function \( \pi_\Omega \) as

\[
\pi_\Omega(z) = \arg\max_{p \in \Delta^d} (p^T z - \Omega(p))
\] (12)

B.2 Characterizing the \( \alpha \)-entmax mapping

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

\[
\alpha\text{-entmax}(z) = [(\alpha - 1)z - \tau^*1]^\frac{1}{\alpha-1}. 
\] (13)

Proof: From the definition of \( \alpha\)-entmax,

\[
\alpha\text{-entmax}(z) := \arg\max_{p \in \Delta^d} p^T z + H^T_\alpha(p),
\] (14)

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

\[
\alpha\text{-entmax}(z) \equiv \pi_{-H^T_\alpha}(z).
\]

We first note that for all \( p \in \Delta^d \),

\[-(\alpha - 1)H^T_\alpha(p) = \frac{1}{\alpha} \sum_{i=1}^d p_i^\alpha + \text{const}. \] (15)

From the constant invariance and scaling properties of \( \pi_\Omega \) (Blondel et al., 2019, Proposition 1, items 4–5),

\[
\pi_{-H^T_\alpha}(z) = \pi_\Omega((\alpha - 1)z), \quad \text{with} \quad \Omega(p) = \sum_{j=1}^d g(p_j), \quad g(t) = \frac{t^\alpha}{\alpha}.
\]

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

\[
\pi_\Omega(z) = [z - \tau^*1]^\frac{1}{\alpha-1}, \quad \text{and therefore} \quad \alpha\text{-entmax}(z) = [(\alpha - 1)z - \tau^*1]^\frac{1}{\alpha-1}. \] (16)

Since \( H^T_\alpha \) is strictly convex on the simplex, \( \alpha\)-entmax has a unique solution \( p^* \). Equation 16 implicitly defines a one-to-one mapping between \( p^* \) and \( \tau^* \) as long as \( p^* \in \Delta \), therefore \( \tau^* \) 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

\[
\text{sparsemax}(z) := \arg\min_{p \in \Delta} \lVert p - z \rVert_2^2. \] (17)

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

\[
\text{sparsemax}(z) = [z - \tau 1]_+. \] (18)
Thus, each coordinate of the sparsemax solution is a piecewise-linear function. Visibly, this expression is recovered when setting $\alpha = 2$ in the $\alpha$-entmax expression (Equation 21); for other values of $\alpha$, the exponent induces curvature.

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

$$\text{softmax}(z)_i := \frac{\exp(z_i)}{\sum_j \exp(z_j)},$$

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

$$\text{softmax}(z)_i = \arg\max_{p \in \Delta} p^\top z + H^S(p),$$

where $H^S(p) := -\sum_i p_i \log p_i$ is the Shannon entropy. Indeed, setting the gradient to 0 yields the condition

$$\log p_i = z_j - \tau^* - 1,$$

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

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

For general values of $\alpha$, Eq. 24 lacks a closed form solution. This makes the computation of the Jacobian 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 $(\alpha, z)$ pairs. We begin

C Jacobian of $\alpha$-entmax w.r.t. the shape parameter $\alpha$: Proof of Proposition 1

Recall that the entmax transformation is defined as:

$$\alpha\text{-entmax}(z) := \arg\max_{p \in \Delta^d} p^\top z + H^T_\alpha(p),$$

where $\alpha \geq 1$ and $H^T_\alpha$ is the Tsallis entropy,

$$H^T_\alpha(p) := \begin{cases} \frac{1}{\alpha-1} \sum_j (p_j - p_j^\alpha), & \alpha \neq 1, \\ H^S(p), & \alpha = 1, \end{cases}$$

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

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

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

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

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

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

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

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

$$\frac{\partial \alpha\text{-entmax}(z)}{\partial \alpha}$$

non-trivial. Fortunately, we can use the technique of implicit differentiation to obtain this Jacobian.
by noting that $\frac{\partial p_i^*}{\partial \alpha} = 0$ if $p_i^* = 0$, because increasing $\alpha$ keeps sparse coordinates sparse.\(^4\) Therefore we need to worry only about coordinates that are in the support of $p^*$. We will assume hereafter that the $i^{th}$ coordinate of $p^*$ is non-zero. We have:

$$\frac{\partial p_i^*}{\partial \alpha} = \frac{\partial}{\partial \alpha} \left[ \frac{1}{\alpha} \log[(\alpha - 1)(z_i - \tau^*)] \right] = p_i^* \frac{\partial}{\partial \alpha} \exp \left[ \frac{1}{\alpha} \log[(\alpha - 1)(z_i - \tau^*)] \right] = \frac{p_i^*}{\alpha - 1} \left[ \frac{\partial}{\partial \alpha} \frac{1}{\alpha} \log[(\alpha - 1)(z_i - \tau^*)] \right] = \frac{p_i^*}{\alpha - 1} \left[ \frac{z_i - \tau^* - (\alpha - 1) \frac{\partial \tau^*}{\partial \alpha}}{z_i - \tau^*} \right] = \frac{p_i^*}{\alpha - 1} \left[ 1 - \frac{\alpha - 1}{z_i - \tau^*} \frac{\partial \tau^*}{\partial \alpha} \right].$$ (26)

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

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

$$0 = \sum_{i \in S} \frac{\partial}{\partial \alpha} \left[ \frac{1}{\alpha} \log[(\alpha - 1)(z_i - \tau^*)] \right]^{1/(\alpha - 1)} = \sum_{i \in S} \frac{p_i^*}{\alpha - 1} \left[ 1 - \frac{\alpha - 1}{z_i - \tau^*} \frac{\partial \tau^*}{\partial \alpha} \right] \log[(\alpha - 1)(z_i - \tau^*)] \right] = \frac{1}{\alpha - 1} \sum_{i \in S} \frac{p_i^*}{(\alpha - 1)(z_i - \tau^*)} - \sum_{i \in S} \frac{p_i^*}{(\alpha - 1)^2} \log[(\alpha - 1)(z_i - \tau^*)] = \frac{1}{\alpha - 1} \sum_{i \in S} (p_i^*)^{2-\alpha} - \sum_{i \in S} \frac{p_i^*}{\alpha - 1} \log p_i$$

$$= \frac{1}{\alpha - 1} \sum_{i \in S} (p_i^*)^{2-\alpha} + \frac{H^S(p^*)}{\alpha - 1},$$ (27)

from which we obtain:

$$\frac{\partial \tau^*}{\partial \alpha} = \frac{1}{(\alpha - 1)^2} + \frac{H^S(p^*)}{\alpha - 1} \frac{1}{\sum_{i}(p_i^*)^{2-\alpha}}.$$ (28)

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

$$\frac{\partial p_i^*}{\partial \alpha} = \frac{p_i^*}{(\alpha - 1)^2} \left[ 1 - \frac{1}{(p_i^*)^{(\alpha - 1)}} \frac{\partial \tau^*}{\partial \alpha} - (\alpha - 1) \log p_i^* \right] = \frac{p_i^*}{(\alpha - 1)^2} \left[ 1 - \frac{1}{(p_i^*)^{(\alpha - 1)}} \frac{H^S(p^*)}{\alpha - 1} \frac{1}{\sum_{i}(p_i^*)^{2-\alpha}} - (\alpha - 1) \log p_i^* \right] = \frac{p_i^*}{(\alpha - 1)^2} - \frac{\hat{p}_i(\alpha)}{(\alpha - 1)^2} \frac{p_i^* \log p_i^*}{\alpha - 1} + \frac{\hat{p}_i(\alpha) H^S(p^*)}{\alpha - 1},$$ (29)

\(^4\)This follows from the margin property of $H^*_\alpha$ (Blondel et al., 2019).
where we denote by

\[ \tilde{\mathbf{p}}(\alpha) = \frac{(p^*_i)^{2-\alpha}}{\sum_j (p^*_j)^{2-\alpha}}. \]  

(30)

The distribution \( \tilde{\mathbf{p}}(\alpha) \) can be interpreted as a “skewed” distribution obtained from \( \mathbf{p}^* \), which appears in the Jacobian of \( \alpha\text{-entmax}(z) \) w.r.t. \( z \) as well (Peters et al., 2019).

C.2 Solving the indetermination for \( \alpha = 1 \)

We can write Eq. 29 as

\[ \frac{\partial p^*_i}{\partial \alpha} = \frac{p^*_i - \tilde{\mathbf{p}}(\alpha) - (\alpha - 1)(p^*_i \log p^*_i + \tilde{\mathbf{p}}(\alpha) H^S(\mathbf{p}^*))}{(\alpha - 1)^2}. \]  

(31)

When \( \alpha \to 1^+ \), we have \( \tilde{\mathbf{p}}(\alpha) \to \mathbf{p}^* \), 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 \( \tilde{\mathbf{p}}_i(\alpha) \) with respect to \( \alpha \). We have

\[ \frac{\partial}{\partial \alpha} (p^*_i)^{2-\alpha} = -(p^*_i)^{2-\alpha} \log p^*_i, \]  

(32)

therefore

\[ \frac{\partial}{\partial \alpha} \tilde{\mathbf{p}}_i(\alpha) = \frac{\partial}{\partial \alpha} \frac{(p^*_i)^{2-\alpha}}{\sum_j (p^*_j)^{2-\alpha}} = -\frac{(p^*_i)^{2-\alpha} \log p^*_i \sum_j (p^*_j)^{2-\alpha} + (p^*_i)^{2-\alpha} \sum_j (p^*_j)^{2-\alpha} \log p^*_j}{(\sum_j (p^*_j)^{2-\alpha})^2} = -\tilde{\mathbf{p}}_i(\alpha) \log p^*_i + \tilde{\mathbf{p}}_i(\alpha) \sum_j \tilde{\mathbf{p}}_j(\alpha) \log p^*_j. \]  

(33)

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

\[ \frac{\partial p^*_i}{\partial \alpha} = \lim_{\alpha \to 1^+} \frac{(1 + (\alpha - 1)H^S(\mathbf{p}^*)) \tilde{\mathbf{p}}_i(\alpha) (\log p^*_i - \sum_j \tilde{\mathbf{p}}_j(\alpha) \log p^*_j) - p^*_i \log p^*_i - \tilde{\mathbf{p}}_i(\alpha) H^S(\mathbf{p}^*)}{2(\alpha - 1)} = A + B, \]  

(34)

with

\[ A = \lim_{\alpha \to 1^+} \frac{H^S(\mathbf{p}^*) \tilde{\mathbf{p}}_i(\alpha) (\log p^*_i - \sum_j \tilde{\mathbf{p}}_j(\alpha) \log p^*_j) H^S(\mathbf{p}^*)}{2} = \frac{H^S(\mathbf{p}^*) p^*_i \log p^*_i + p^*_i (H^S(\mathbf{p}^*))^2}{2}, \]  

(35)

and

\[ B = \lim_{\alpha \to 1^+} \frac{\tilde{\mathbf{p}}_i(\alpha) (\log p^*_i - \sum_j \tilde{\mathbf{p}}_j(\alpha) \log p^*_j) - p^*_i \log p^*_i - \tilde{\mathbf{p}}_i(\alpha) H^S(\mathbf{p}^*)}{2(\alpha - 1)}. \]  

(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:

\[
B = \frac{1}{2} \lim_{\alpha \to 1^+} \left\{ \tilde{p}_i(\alpha) \log p_i^* \left( \sum_j \tilde{p}_j(\alpha) \log p_j^* - \log p_i^* \right) 
- \tilde{p}_i(\alpha) \left( \sum_j \tilde{p}_j(\alpha) \log p_j^* - \log p_i^* \right) \left( \sum_j \tilde{p}_j(\alpha) \log p_j^* + H^S(p^*) \right) 
- \tilde{p}_i(\alpha) \sum_j \tilde{p}_j(\alpha) \log p_j^* \left( \sum_k \tilde{p}_k(\alpha) \log p_k^* - \log p_j^* \right) \right\} 
- \frac{p_i^* \log p_i^*(H^S(p^*) + \log p_i^*) + p_i^* \sum_j p_j^* \log p_j^*(H^S(p^*) + \log p_j^*)}{2} 
+ \frac{-H^S(p^*)p_i^* \log p_i^* - p_i^*(H^S(p^*))^2 - p_i^* \log^2 p_i^* + p_i^* \sum_j p_j^* \log^2 p_j^*}{2}. \tag{37}
\]

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

\[
\frac{\partial p_i^*}{\partial \alpha} \bigg|_{\alpha=1} = \frac{-p_i^* \log^2 p_i^* + p_i^* \sum_j p_j^* \log^2 p_j^*}{2}. \tag{38}
\]

C.3 Summary

To sum up, we have the following expression for the Jacobian of \( \alpha \)-entmax with respect to \( \alpha \):

\[
\frac{\partial p_i^*}{\partial \alpha} = \begin{cases} 
\frac{p_i^* - \tilde{p}_i(\alpha)}{(\alpha - 1)^2} - \frac{p_i^* \log p_i^* + H^S(p^*)}{\alpha - 1}, & \text{for } \alpha > 1 \\
-\frac{p_i^* \log^2 p_i^* + p_i^* \sum_j p_j^* \log^2 p_j^*}{2}, & \text{for } \alpha = 1.
\end{cases} \tag{39}
\]