A. Supplementary Material

A.1. Proof of Prop. 1

The Lagrangian of the optimization problem in Eq. 2 is:

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}, \tau) = \frac{1}{2}\|\boldsymbol{p} - \boldsymbol{z}\|^2 - \boldsymbol{\mu}^\top \boldsymbol{p} + \tau(\mathbf{1}^\top \boldsymbol{p} - 1). \tag{37}$$

The optimal $(\boldsymbol{p}^*, \boldsymbol{\mu}^*, \tau^*)$ must satisfy the following Karush-Kuhn-Tucker conditions:

$$\boldsymbol{p}^* - \boldsymbol{z} - \boldsymbol{\mu}^* + \tau^* \mathbf{1} = \mathbf{0}, \tag{38}$$

$$\mathbf{1}^\top \boldsymbol{p}^* = 1, \quad \boldsymbol{p}^* \geq \mathbf{0}, \quad \boldsymbol{\mu}^* \geq \mathbf{0}, \tag{39}$$

$$\mu_i^* p_i^* = 0, \ \forall i \in [K]. \tag{40}$$

If for $i \in [K]$ we have $p_i^* > 0$, then from Eq. 40 we must have $\mu_i^* = 0$, which from Eq. 38 implies $p_i^* = z_i - \tau^*$. Let $S(\boldsymbol{z}) = \{j \in [K] \mid p_j^* > 0\}$. From Eq. 39 we obtain $\sum_{j \in S(\boldsymbol{z})}(z_j - \tau^*) = 1$, which yields the right hand side of Eq. 4. Again from Eq. 40, we have that $\mu_i^* > 0$ implies $p_i^* = 0$, which from Eq. 38 implies $\mu_i^* = \tau^* - z_i \geq 0$, i.e., $z_i \leq \tau^*$ for $i \notin S(\boldsymbol{z})$. Therefore we have that $k(\boldsymbol{z}) = |S(\boldsymbol{z})|$, which proves the first equality of Eq. 4.

A.2. Proof of Prop. 2

We start with the third property, which follows from the coordinate-symmetry in the definitions in Eqs. 1–2. The same argument can be used to prove the first part of the first property (uniform distribution).

Let us turn to the second part of the first property (peaked distribution on the maximal components of $\boldsymbol{z}$), and define $t = \epsilon^{-1}$. For the softmax case, this follows from

$$\lim_{t \to +\infty} \frac{e^{tz_i}}{\sum_k e^{tz_k}} = \lim_{t \to +\infty} \frac{e^{tz_i}}{\sum_{k \in A(\boldsymbol{z})} e^{tz_k}} = \lim_{t \to +\infty} \frac{e^{t(z_i - z_{(1)})}}{|A(\boldsymbol{z})|} = \begin{cases} 1/|A(\boldsymbol{z})|, & \text{if } i \in A(\boldsymbol{z}) \\ 0, & \text{otherwise.} \end{cases} \tag{41}$$

For the sparsemax case, we invoke Eq. 4 and the fact that $k(t\boldsymbol{z}) = |A(\boldsymbol{z})|$ if $\gamma(t\boldsymbol{z}) \geq 1/|A(\boldsymbol{z})|$. Since $\gamma(t\boldsymbol{z}) = t\gamma(\boldsymbol{z})$, the result follows.

The second property holds for softmax, since $(e^{z_i + c}) / \sum_k e^{z_k + c} = e^{z_i} / \sum_k e^{z_k}$; and for sparsemax, since for any $\boldsymbol{p} \in \Delta^{K-1}$ we have $\|\boldsymbol{p} - \boldsymbol{z} - c\mathbf{1}\|^2 = \|\boldsymbol{p} - \boldsymbol{z}\|^2 - 2c\mathbf{1}^\top(\boldsymbol{p} - \boldsymbol{z}) + \|c\mathbf{1}\|^2$, which equals $\|\boldsymbol{p} - \boldsymbol{z}\|^2$ plus a constant (because $\mathbf{1}^\top \boldsymbol{p} = 1$).

Finally, let us turn to fourth property. The first inequality states that $z_i \leq z_j \ \Rightarrow \ \rho_i(\boldsymbol{z}) \leq \rho_j(\boldsymbol{z})$ (*i.e.*, coordinate monotonicity). For the softmax case, this follows trivially from the fact that the exponential function is increasing. For the sparsemax, we use a proof by contradiction. Suppose $z_i \leq z_j$ and $\text{sparsemax}_i(\boldsymbol{z}) > \text{sparsemax}_j(\boldsymbol{z})$. From the definition in Eq. 2, we must have $\|\boldsymbol{p} - \boldsymbol{z}\|^2 \geq \|\text{sparsemax}(\boldsymbol{z}) - \boldsymbol{z}\|^2$, for any $\boldsymbol{p} \in \Delta^{K-1}$. This leads to a contradiction if we choose $p_k = \text{sparsemax}_k(\boldsymbol{z})$ for $k \notin \{i, j\}$, $p_i = \text{sparsemax}_j(\boldsymbol{z})$, and $p_j = \text{sparsemax}_i(\boldsymbol{z})$. To prove the second inequality in the fourth property for softmax, we need to show that, with $z_i \leq z_j$, we have $(e^{z_j} - e^{z_i}) / \sum_k e^{z_k} \leq (z_j - z_i)/2$. Since $\sum_k e^{z_k} \geq e^{z_j} + e^{z_i}$, it suffices to consider the binary case, *i.e.*, we need to prove that $\tanh((z_j - z_i)/2) = (e^{z_j} - e^{z_i})/(e^{z_j} + e^{z_i}) \leq (z_j - z_i)/2$, that is, $\tanh(t) \leq t$ for $t \geq 0$. This comes from $\tanh(0) = 0$ and $\tanh'(t) = 1 - \tanh^2(t) \leq 1$. For sparsemax, given two coordinates i, j, three things can happen: (i) both are thresholded, in which case $\rho_j(\boldsymbol{z}) - \rho_i(\boldsymbol{z}) = z_j - z_i$; (ii) the smaller ($z_i$) is truncated, in which case $\rho_j(\boldsymbol{z}) - \rho_i(\boldsymbol{z}) = z_j - \tau(\boldsymbol{z}) \leq z_j - z_i$; (iii) both are truncated, in which case $\rho_j(\boldsymbol{z}) - \rho_i(\boldsymbol{z}) = 0 \leq z_j - z_i$.

A.3. Proof of Prop. 3

To prove the first claim, note that, for $j \in S(\boldsymbol{z})$,

$$\frac{\partial \tau^2(\boldsymbol{z})}{\partial z_j} = 2\tau(\boldsymbol{z}) \frac{\partial \tau(\boldsymbol{z})}{\partial z_j} = \frac{2\tau(\boldsymbol{z})}{|S(\boldsymbol{z})|}, \tag{42}$$

where we used Eq. 10. We then have

$$\frac{\partial L_{\text{sparsemax}}(\boldsymbol{z}; k)}{\partial z_j} = \begin{cases} -\delta_k(j) + z_j - \tau(\boldsymbol{z}) & \text{if } j \in S(\boldsymbol{z}) \\ -\delta_k(j) & \text{otherwise.} \end{cases}$$

That is, $\nabla_{\boldsymbol{z}} L_{\mathrm{sparsemax}}(\boldsymbol{z}; k) = -\boldsymbol{\delta}_k + \mathrm{sparsemax}(\boldsymbol{z})$.

To prove the second statement, from the expression for the Jacobian in Eq. 11, we have that the Hessian of $L_{\mathrm{sparsemax}}$ (strictly speaking, a "sub-Hessian" (Penot, 2014), since the loss is not twice-differentiable everywhere) is given by

$$\frac{\partial^2 L_{\mathrm{sparsemax}}(\boldsymbol{z}; k)}{\partial x_i \partial x_j} = \begin{cases} \delta_{ij} - \frac{1}{|S(\boldsymbol{z})|} & \text{if } i, j \in S(\boldsymbol{z}) \\ 0 & \text{otherwise.} \end{cases} \tag{43}$$

This Hessian can be written in the form $\mathrm{Id} - \mathbf{1}\mathbf{1}^\top / |S(\boldsymbol{z})|$ up to padding zeros (for the coordinates not in $S(\boldsymbol{z})$), where Id is the identity matrix; hence it is positive semi-definite (with rank $|S(\boldsymbol{z})| - 1$), which establishes the convexity of $L_{\mathrm{sparsemax}}$.

For the third claim, we have $L_{\mathrm{sparsemax}}(\boldsymbol{z} + c\mathbf{1}) = -z_k - c + \frac{1}{2}\sum_{j \in S(\boldsymbol{z})}(z_j^2 - \tau^2(\boldsymbol{z}) + 2c(z_j - \tau)) + \frac{1}{2} = -z_k - c + \frac{1}{2}\sum_{j \in S(\boldsymbol{z})}(z_j^2 - \tau^2(\boldsymbol{z}) + 2cp_j) + \frac{1}{2} = L_{\mathrm{sparsemax}}(\boldsymbol{z})$, since $\sum_{j \in S(\boldsymbol{z})} p_j = 1$.

From the first two claims, we have that the minima of $L_{\mathrm{sparsemax}}$ have zero gradient, i.e., satisfy the equation $\mathrm{sparsemax}(\boldsymbol{z}) = \boldsymbol{\delta}_k$. Furthemore, from Prop. 2, we have that the sparsemax never increases the distance between two coordinates, i.e., $\mathrm{sparsemax}_k(\boldsymbol{z}) - \mathrm{sparsemax}_j(\boldsymbol{z}) \leq z_k - z_j$. Therefore $\mathrm{sparsemax}(\boldsymbol{z}) = \boldsymbol{\delta}_k$ implies $z_k \geq 1 + \max_{j \neq k} z_j$. To prove the converse statement, note that the distance above can only be decreased if the smallest coordinate is truncated to zero. This establishes the equivalence between (ii) and (iii) in the fifth claim. Finally, we have that the minimum loss value is achieved when $S(\boldsymbol{z}) = \{k\}$, in which case $\tau(\boldsymbol{z}) = z_k - 1$, leading to

$$L_{\mathrm{sparsemax}}(\boldsymbol{z}; k) = -z_k + \frac{1}{2}(z_k^2 - (z_k - 1)^2) + \frac{1}{2} = 0. \tag{44}$$

This proves the equivalence with (i) and also the fourth claim.