# From Softmax to Sparsemax

## A Sparse Model of Attention and Multi-Label Classification

André F. T. Martins and Ramón F. Astudillo

ICML 2016

# Softmax

- ❖ Converting a representation vector into a posterior probabilities of labels $\mathbb{R}^K \rightarrow \Delta^{K-1}$, defined as:

$$\text{softmax}_i(z) = \frac{\exp\left(z_i\right)}{\sum_{k=1}^{K} \exp\left(z_k\right)}$$

- ❖ Limitation: full support, $\text{softmax}(z) > 0, \forall z$

- ❖ Requires: sparse probability distribution by assign exactly zero probability —> interpretability !

# Sparsemax

- A sparse alternative: Euclidean projection of z onto the probability simplex

$$\text{sparsemax}(z) := \underset{\boldsymbol{p} \in \Delta^{K-1}}{\operatorname{argmin}} \|\boldsymbol{p} - \boldsymbol{z}\|^2$$

- Close-Form Solution

$$\text{sparsemax}_i(z) = \max\left\{0,\ z_i - \tau\right\}$$

$\tau$ is a normalizing threshold function such that $\sum_j \max\left\{0, z_j - \tau\right\} = 1$

- How to compute $\tau$ ?

# Sparsemax Evaluation

---

**Algorithm 1** Sparsemax Evaluation

---

**Input:** $\boldsymbol{z}$

Sort $\boldsymbol{z}$ as $z_{(1)} \geq \ldots \geq z_{(K)}$

Find $k(\boldsymbol{z}) := \max \left\{ k \in [K] \mid 1 + k z_{(k)} > \sum_{j \leq k} z_{(j)} \right\}$

Define $\tau(\boldsymbol{z}) = \frac{\left(\sum_{j \leq k(\boldsymbol{z})} z_{(j)}\right) - 1}{k(\boldsymbol{z})}$

**Output:** $\boldsymbol{p}$ s.t. $p_i = [z_i - \tau(\boldsymbol{z})]_+$.

---

Sort the coordinates of z

Define k(z) the sparsity bound: index

Sparse logits $\{z_{(j)} | j \leq k(\mathbf{z})\}$ after threshold; sum of sparse logits $\sum_{j \leq k(z)} z(j)$

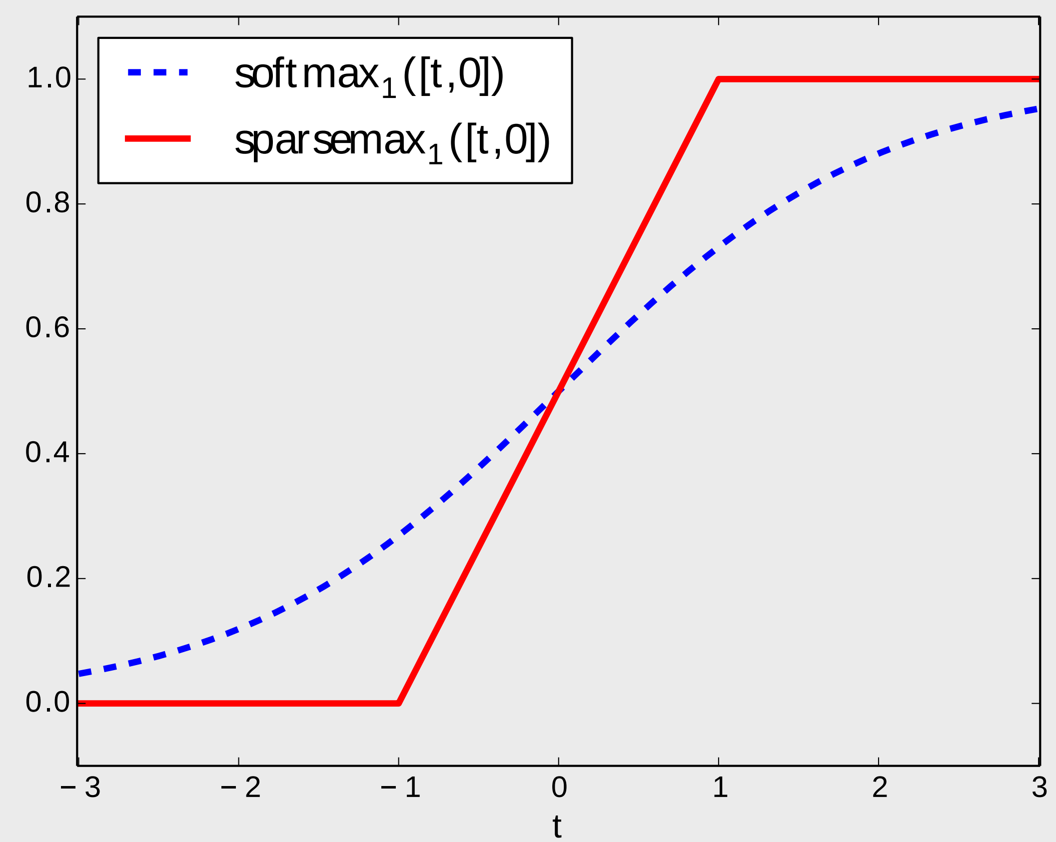
$\tau(z)$ is the threshold function

$S(z) := \left\{ j \in [K] \,|\, \text{sparsemax}_j(z) > 0 \right\}$ the support of sparsemax

# Two and Three-Dimensional Cases

- For two dimensions: $z = (t, 0)$, softmax becomes logistic (sigmoid) function as:

$$\text{softmax}_1(z) = (1 + \exp(-t))^{-1}$$




- 2D sparsemax is the "hard" version of the sigmoid

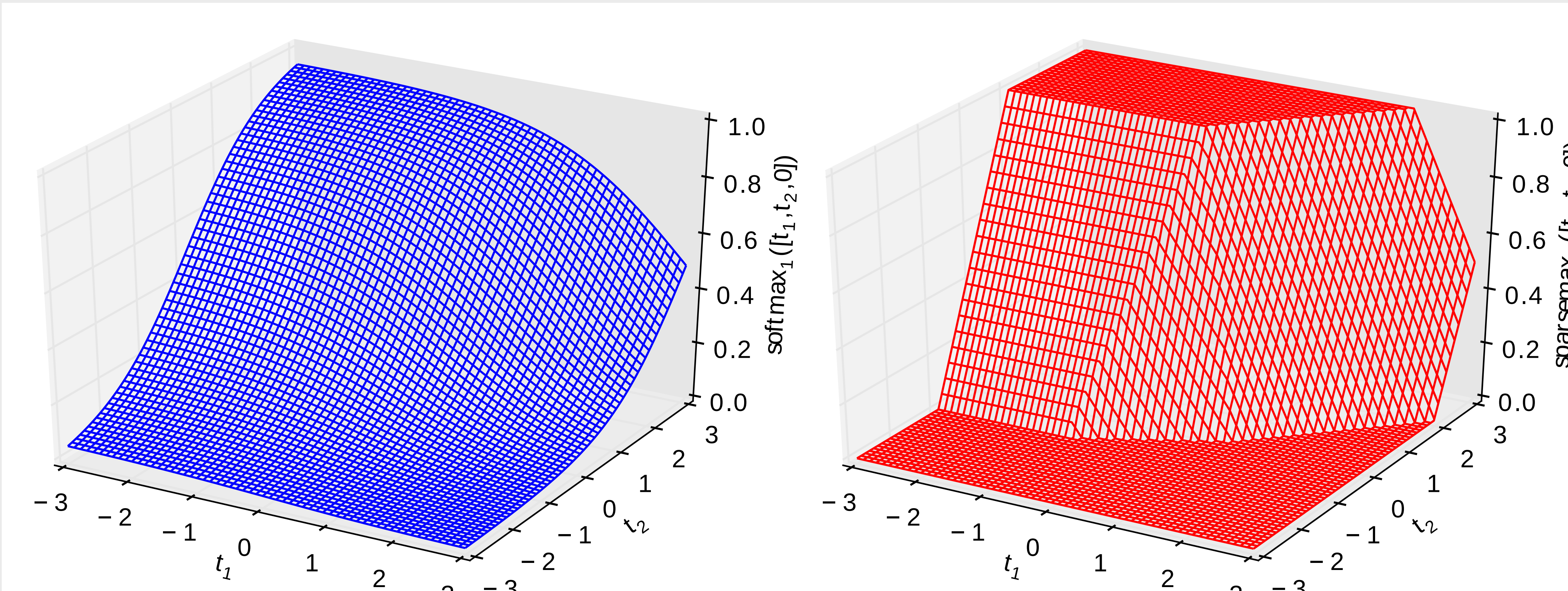
For $z = (t, 0)$

$$\tau(\boldsymbol{z}) = \begin{cases} t-1, & \text{if } t > 1 \\ (t-1)/2, & \text{if } -1 \le t \le 1 \\ -1, & \text{if } t < -1 \end{cases}$$

$$\text{sparsemax}_1(\boldsymbol{z}) = \begin{cases} 1, & \text{if } t > 1 \\ (t+1)/2, & \text{if } -1 \le t \le 1 \\ 0, & \text{if } t < -1 \end{cases}$$

piece-wise linear

# Two and Three-Dimensional Cases




**5D case**

```
Logits
tensor([[ 0.9549,  0.4015,  1.3101,  0.5750, -1.9022],
        [-1.7090, -0.5747, -0.1654,  0.1718,  0.1057]], device='cuda:0')

Softmax probabilities
tensor([[0.2672, 0.1536, 0.3811, 0.1827, 0.0153],
        [0.0465, 0.1447, 0.2179, 0.3052, 0.2857]], device='cuda:0')

Sparsemax probabilities
tensor([[0.3224, 0.0000, 0.6776, 0.0000, 0.0000],
        [0.0000, 0.0000, 0.1305, 0.4678, 0.4017]], device='cuda:0')
```

# Gradient-based Optimization

❖ Jacobian of Softmax

$$\frac{\partial\, \mathrm{softmax}_i(\boldsymbol{z})}{\partial z_j} = \frac{\delta_{ij} e^{z_i} \sum_k e^{z_k} - e^{z_i} e^{z_j}}{\left(\sum_k e^{z_k}\right)^2} = \mathrm{softmax}_i(\boldsymbol{z})\left(\delta_{ij} - \mathrm{softmax}_j(\boldsymbol{z})\right)$$

For matrix notation with $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$,

$$\mathbf{J}_{\mathrm{softmax}}(\boldsymbol{z}) = \mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$$

# Gradient-based Optimization

❖ Jacobian of Sparsemax

1. $$\frac{\partial \text{sparsemax}_i(\boldsymbol{z})}{\partial z_j} = \begin{cases} \delta_{ij} - \frac{\partial \tau(\boldsymbol{z})}{\partial z_j}, & \text{if } z_i > \tau(\boldsymbol{z}) \\ 0, & \text{if } z_i \leq \tau(\boldsymbol{z}) \end{cases}$$

2. $$\frac{\partial \tau(\boldsymbol{z})}{\partial z_j} = \begin{cases} \frac{1}{|S(\boldsymbol{z})|} & \text{if } j \in S(\boldsymbol{z}) \\ 0, & \text{if } j \notin S(\boldsymbol{z}) \end{cases}$$

3. $$\frac{\partial \text{sparsemax}_i(\boldsymbol{z})}{\partial z_j} = \begin{cases} \delta_{ij} - \frac{1}{|S(\boldsymbol{z})|}, & \text{if } i, j \in S(\boldsymbol{z}) \\ 0, & \text{otherwise} \end{cases}$$

4. $\mathbf{J}_{\text{sparsemax}}(\boldsymbol{z}) = \mathbf{D}\,\text{iag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top / |S(\boldsymbol{z})|$ where $\boldsymbol{s} = \text{sparsemax}(\boldsymbol{z})$

$$\mathbf{sparsemax}_i(z) = \max\left\{0,\ z_i - \tau\right\}$$

$$\tau(z) = \frac{\left(\sum_{j \leq k(z)} z_{(j)}\right) - 1}{k(z)}$$

$$S(z) := \left\{ j \in [K] \,|\, \text{sparsemax}_j(z) > 0 \right\}$$

# Loss Function: Logistic Loss

❖ Consider regularized empirical risk minimization problems

$$\text{minimize } \frac{\lambda}{2}\|\mathbf{W}\|_F^2 + \frac{1}{N}\sum_{i=1}^{N} L\left(\mathbf{W}\boldsymbol{x}_i + \boldsymbol{b}; y_i\right)$$

❖ Logistic loss

$$L_{\text{softmax}}(z; k) = -\log \text{softmax}_k(z) = -z_k + \log \sum_j \exp\left(z_j\right)$$

❖ Gradient

$$\nabla_{\boldsymbol{z}} L_{\text{softmax}}(\boldsymbol{z}; k) = -\boldsymbol{\delta}_k + \text{softmax}(\boldsymbol{z})$$

$\boldsymbol{\delta}_k$ is the delta distribution on k.

# Loss Function: Sparsemax Loss

❖ Reversing engineering the sparsemax loss

$$\nabla_{\boldsymbol{z}} L_{\text{sparsemax}}(\boldsymbol{z}; k) = -\boldsymbol{\delta}_k + \text{sparsemax}(\boldsymbol{z})$$

❖ Sparsemax loss:

$$L_{\text{sparsemax}}(z; k) = -z_k + \frac{1}{2} \sum_{j \in S(z)} \left( z_j^2 - \tau^2(z) \right) + \frac{1}{2}$$

# Generalization to Multi-Label Classification

❖ The multinomial logistic loss

$$L_{\text{softmax}}(\boldsymbol{z};\boldsymbol{q}) = \boldsymbol{KL}(\boldsymbol{q} \| \operatorname{softmax}(\boldsymbol{z})) = -\mathbf{H}(\boldsymbol{q}) - \boldsymbol{q}^\top \boldsymbol{z} + \log \sum_j \exp\left(z_j\right)$$

Gradient: $\nabla_z L_{\text{softmax}}(\boldsymbol{z};\boldsymbol{q}) = -\boldsymbol{q} + \operatorname{softmax}(\boldsymbol{z})$

❖ The corresponding generalization in the sparsemax case

$$L_{\text{sparsemax}}(\boldsymbol{z};\boldsymbol{q}) = -\boldsymbol{q}^\top \boldsymbol{z} + \frac{1}{2} \sum_{j \in S(z)} \left(z_j^2 - \tau^2(\boldsymbol{z})\right) + \frac{1}{2}\|\boldsymbol{q}\|^2$$

Gradient: $\nabla_z L_{\text{sparsemax}}(\boldsymbol{z};\boldsymbol{q}) = -\boldsymbol{q} + \operatorname{sparsemax}(\boldsymbol{z})$

# Experiments: Multi-Label Benchmarking

*Table 1.* Statistics for the 5 multi-label classification datasets.

| DATASET | DESCR. | #LABELS | #TRAIN | #TEST |
|---|---|---|---|---|
| SCENE | IMAGES | 6 | 1211 | 1196 |
| EMOTIONS | MUSIC | 6 | 393 | 202 |
| BIRDS | AUDIO | 19 | 323 | 322 |
| CAL500 | MUSIC | 174 | 400 | 100 |
| REUTERS | TEXT | 103 | 23,149 | 781,265 |

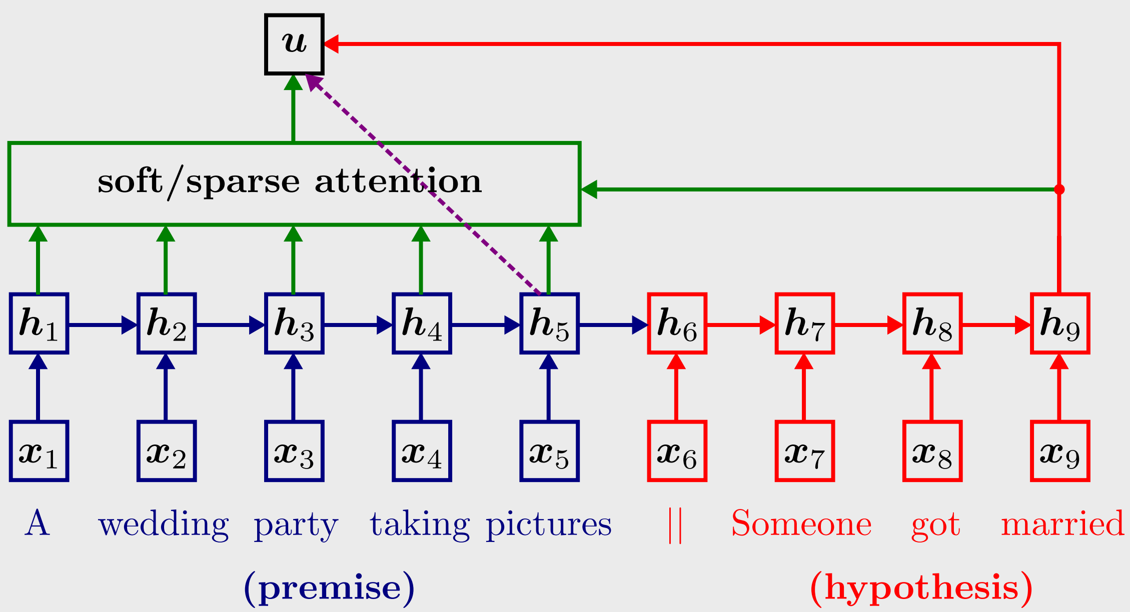
*Table 2.* Micro (left) and macro-averaged (right) $F_1$ scores for the logistic, softmax, and sparsemax losses on benchmark datasets.

| DATASET | LOGISTIC | SOFTMAX | SPARSEMAX |
|---|---|---|---|
| SCENE | 70.96 / 72.95 | **74.01** / **75.03** | 73.45 / 74.57 |
| EMOTIONS | 66.75 / **68.56** | **67.34** / 67.51 | 66.38 / 66.07 |
| BIRDS | 45.78 / 33.77 | 48.67 / 37.06 | **49.44** / **39.13** |
| CAL500 | **48.88** / 24.49 | 47.46 / 23.51 | 48.47 / **26.20** |
| REUTERS | **81.19** / 60.02 | 79.47 / 56.30 | 80.00 / **61.27** |

- Logistic: independent binary logistic regressors on each label
- Softmax: a multinomial logistic regressor
- a slight advantage of Sparsemax

# Experiments:
# Neural Networks with Attention Mechanisms

## ❖ Sparse Attention




| | Dev Acc. | Test Acc. |
|---|---|---|
| NoAttention | 81.84 | 80.99 |
| LogisticAttention | 82.11 | 80.84 |
| SoftAttention | 82.86 | 82.08 |
| SparseAttention | 82.52 | **82.20** |

**NoAttention:**

$$\boldsymbol{u} = \tanh\left(\mathbf{W}^{pu}\boldsymbol{h}_L + \mathbf{W}^{hu}\boldsymbol{h}_N + \boldsymbol{b}^u\right)$$

**SoftAttention:**

$$z_t = \boldsymbol{v}^\top \tanh\left(\mathbf{W}^{pm}\boldsymbol{h}_t + \mathbf{W}^{hm}\boldsymbol{h}_N + \boldsymbol{b}^m\right)$$

$$\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z}),\ \ \textbf{where}\ \boldsymbol{z} := \left(z_1, \ldots, z_L\right)$$

---

A boy **rides on** a **camel** in a crowded area while talking on his cellphone.
Hypothesis: *A boy is riding an animal.* [entailment]

---

A young girl wearing **a pink coat** plays with a **yellow** toy golf club.
Hypothesis: *A girl is wearing a blue jacket.* [contradiction]

---

Two black dogs are **frolicking** around the **grass together**.
Hypothesis: *Two dogs swim in the lake.* [contradiction]

---

A man wearing a yellow striped shirt **laughs** while **seated next** to another **man** who is wearing a light blue shirt and **clasping** his hands together.
Hypothesis: *Two mimes sit in complete silence.* [contradiction]

---