NeurIPS | 2018

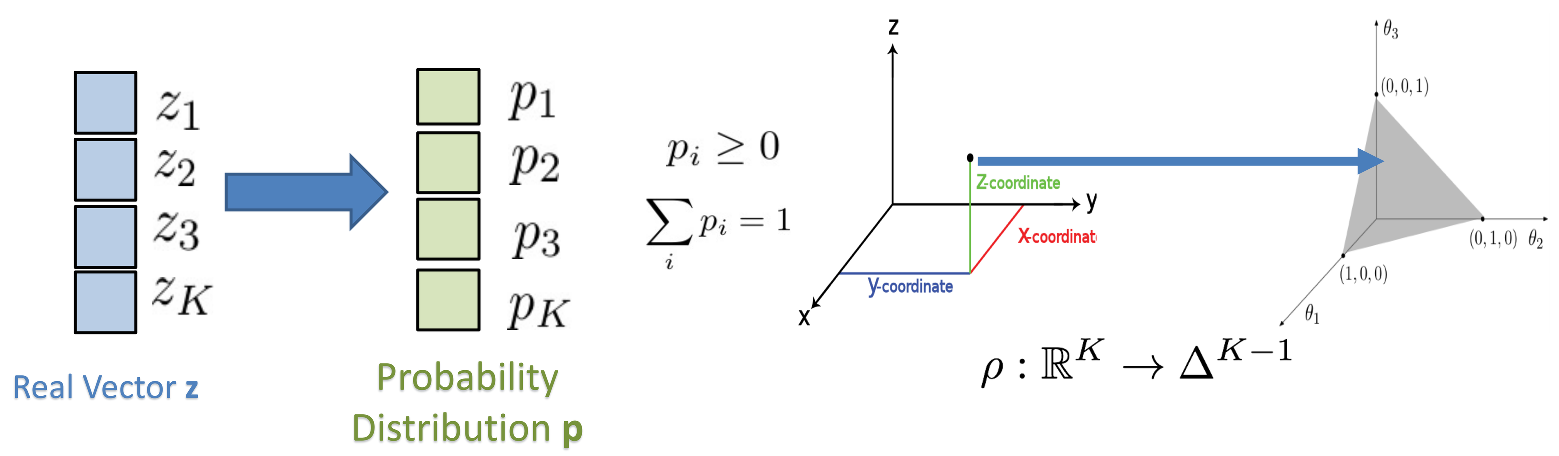
On Controllable Sparse Alternatives to Softmax

***Anirban Laha*[1]*, *Saneem A. Chemmengath*[1]*, ***Priyanka Agrawal*[1]**, *Mitesh M. Khapra*[2], *Karthik Sankaranarayanan*[1], *Harish G. Ramaswamy*[2]

[1]{**anirlaha**, saneem.cg, **priyanka.agrawal**, kartsank}@in.ibm.com (**IBM Research, India**), [2]{miteshk, hariguru}@cse.iitm.ac.in (**IIT Madras**)

Probability Mapping Functions

- We are looking for a function $\boldsymbol{\rho}$ which takes a real vector **z** and produces a probability distribution **p.**



- Applications:
  1. **Probabilistic classification –** Multiclass classification, Multilabel Classification.
  2. **Neural Attention Models –** Attention networks need a probability distribution over input states while generating output states.
  3. **Memory Networks, Reinforcement Learning, Knowledge distillation** and many more.

- Known probability mapping functions:
  1. Softmax $\rho_i(\boldsymbol{z}) = \frac{\exp(z_i)}{\sum_{j\in[K]}\exp(z_j)}$
  2. Sum-normalization $\rho_i(\boldsymbol{z}) = \frac{z_i}{\sum_{j\in[K]} z_j}$
  3. Spherical softmax $\rho_i(\boldsymbol{z}) = \frac{z_i^2}{\sum_{j\in[K]} z_j^2}$

Limitations:
1. Cannot be sparse
2. Not full-domain
3. Not monotonic

- Need for sparsity
  1. In multilabel classification, ONLY A FEW labels out of 1000s of possible labels are TRUE.
  2. In attention models/memory networks, sparser probabilities make COMPUTATION FASTER.
- Sparsity in output VERSUS sparsity in model parameters.

$$y = \rho(W^\top x + b)$$

Sparsity in output

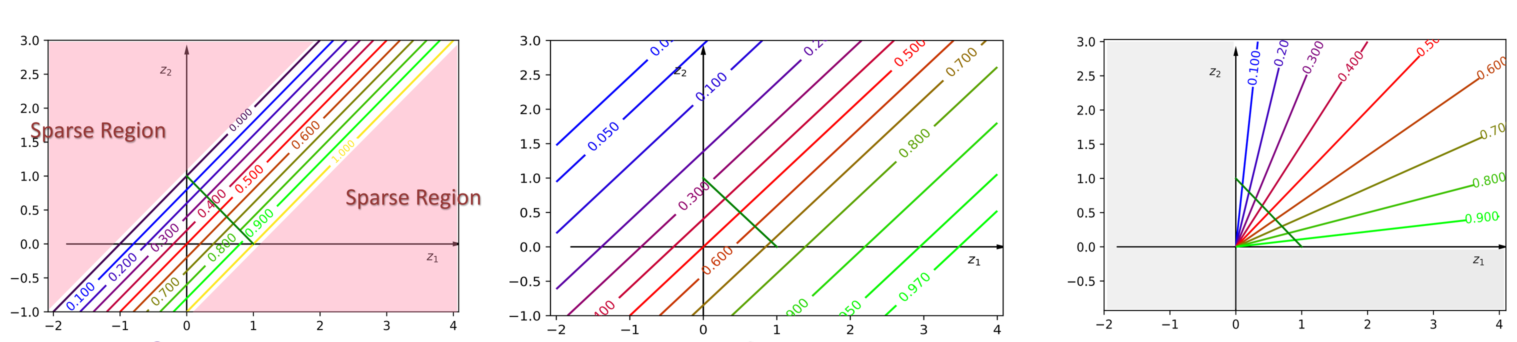
Sparsity in parameter

Sparse Probability Maps

- Sparsemax – Projection onto simplex [ICML 2016]:

$$\rho(\boldsymbol{z}) = \underset{\boldsymbol{p}\in\Delta^{K-1}}{\operatorname{argmin}} \|\boldsymbol{p}-\boldsymbol{z}\|_2^2$$

No control over sparsity!!



Sparsemax | Softmax | Sum-normalization

Sparsegen – Unified Framework

- A family of sparse probability mapping functions:

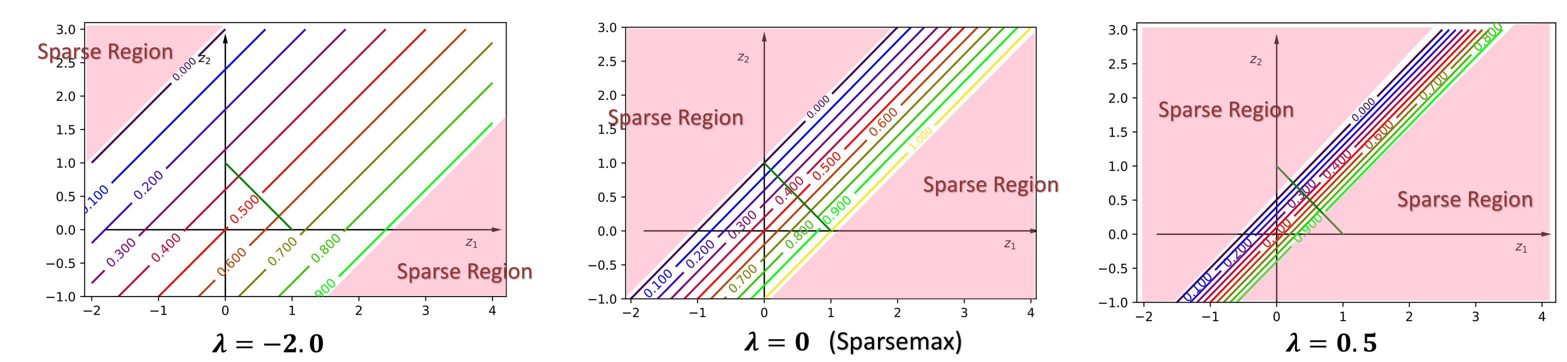
$$\rho(\boldsymbol{z}) = \text{sparsegen}(\boldsymbol{z}; g, \lambda) = \underset{\boldsymbol{p}\in\Delta^{K-1}}{\operatorname{argmin}} \|\boldsymbol{p}-g(\boldsymbol{z})\|_2^2 - \lambda\|\boldsymbol{p}\|_2^2 \qquad g: \mathbb{R}^K \to \mathbb{R}^K$$

- Closed Form solution exists.

Controls for Sparsity

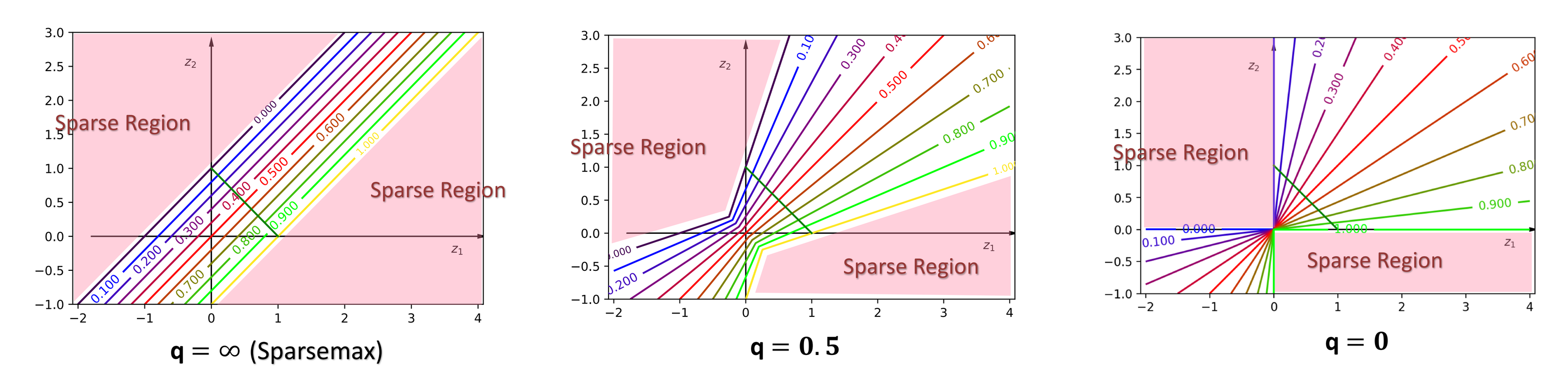
- **Sparsegen-lin** (Control over width of non-sparse region):

$$\rho(\boldsymbol{z}) = \text{sparsegen-lin}(\boldsymbol{z}) = \underset{\boldsymbol{p}\in\Delta^{K-1}}{\operatorname{argmin}} \|\boldsymbol{p}-\boldsymbol{z}\|_2^2 - \lambda\|\boldsymbol{p}\|_2^2$$



- **Sparse-hourglass** (Control over shape of non-sparse region):

$$\rho(\boldsymbol{z}) = \text{sparsehourglass}(\boldsymbol{z}) = \underset{\boldsymbol{p}\in\Delta^{K-1}}{\operatorname{argmin}} \left\|\boldsymbol{p} - \frac{1+Kq}{|\sum_{i\in[K]} z_i| + Kq}\boldsymbol{z}\right\|_2^2$$



Translation Invariance!!! Scale Invariance!!!

- Parameter **q** helps trade-off between translation and scale invariances.
  1. *Translation Invariance*: Adding a constant value to all dimensions of **z** keeps **p** unchanged.
  2. *Scale Invariance*: Multiplying all dimensions of **z** by a constant value keeps **p** unchanged.

Sparsity Inducing Loss Functions

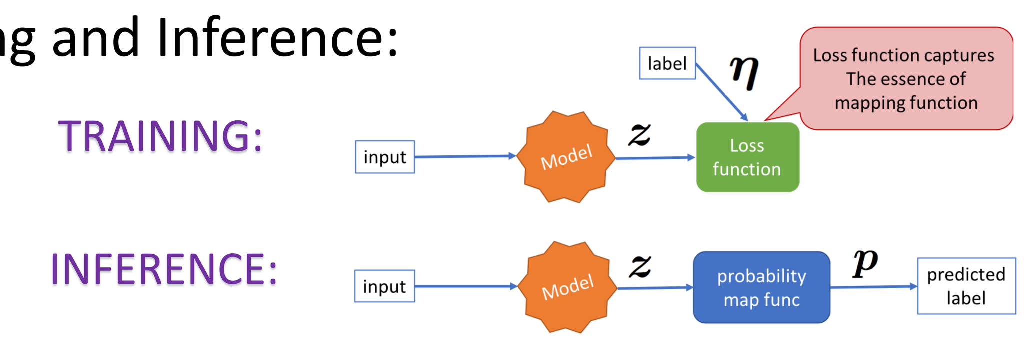
- **Setting: Multilabel Classification**

  -- More than one labels for an instance can be true.

  -- *Usual approach* : Separate logistic sigmoid based binary classifier for every label followed by thresholding.

  -- *In this work*: Apply sparse probability mapping function. Non-zeroes are predicted labels, zeroes are non-labels.

- Training and Inference:



- Convex hinge-based loss functions:

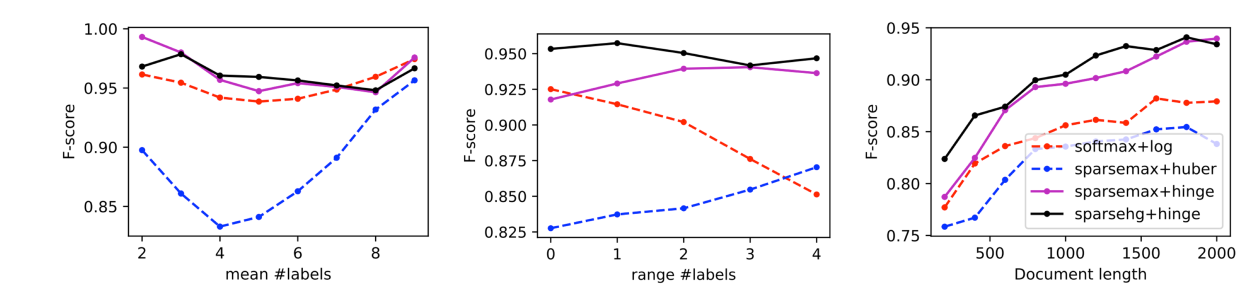
$$\mathcal{L}_{\text{sparsehg,hinge}}(\boldsymbol{z},\boldsymbol{\eta}) = \sum_{\substack{i,j\\ \eta_i\neq 0,\eta_j\neq 0}} |z_i - z_j| + \sum_{\substack{i,j\\ \eta_i\neq 0,\eta_j=0}} \max\left\{\frac{\eta_i}{\hat{\alpha}(\boldsymbol{z})} - (z_i - z_j), 0\right\} \qquad \hat{\alpha}(\boldsymbol{z}) = \frac{1+Kq}{|\sum_i z_i| + Kq}$$

$$\mathcal{L}_{\text{sparsegen-lin,hinge}}(\boldsymbol{z},\boldsymbol{\eta}) = \frac{1}{1-\lambda}\sum_{\substack{i,j\\ \eta_i\neq 0,\eta_j\neq 0}} |z_i - z_j| + \sum_{\substack{i,j\\ \eta_i\neq 0,\eta_j=0}} \max\left\{\eta_i - \frac{z_i - z_j}{1-\lambda}, 0\right\}$$

Multilabel Classification

Synthetic Multilabel Experimental Setup:
a) Varying mean #labels.
b) Varying range #labels.
c) Varying document length.

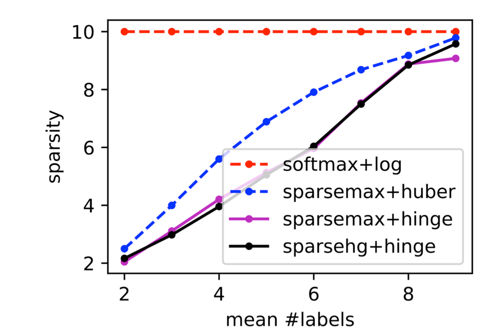
Competing models:
1. Baseline **softmax+log**.
2. Baseline **sparsemax+huber** [ICML 2016].
3. Proposed **sparsemax+hinge**.
4. Proposed **sparsehg+hinge**.



(a) Varying mean #labels (b) Varying range #labels (c) Varying document length

More accurate predictions

Sparser outputs

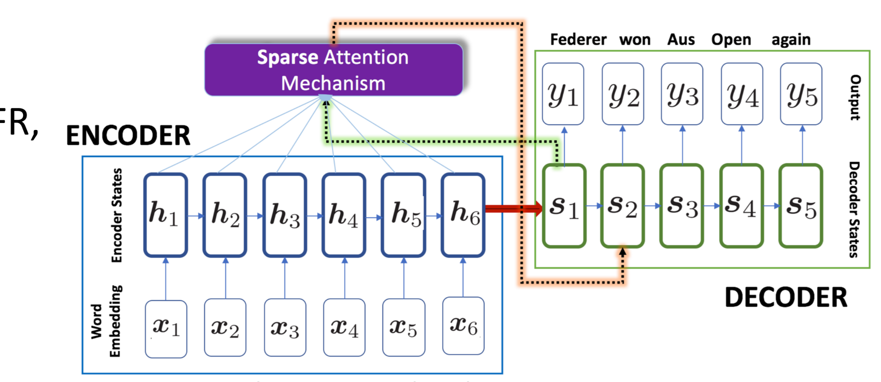


Controlled Sparse Attention

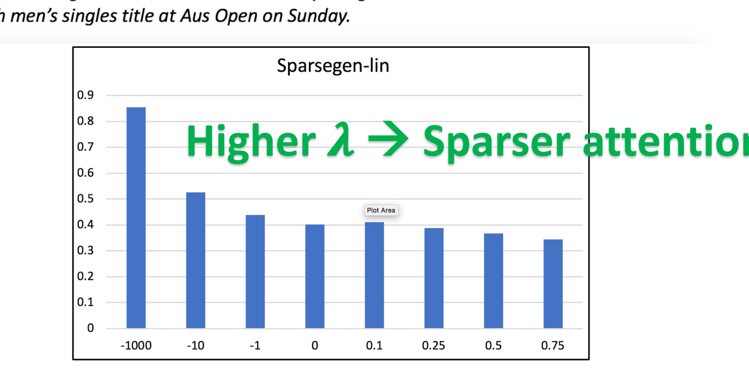
Seq2seq Models with attention:

a) Neural Machine Translation (EN-FR, FR-EN).
b) Abstractive Summarization (Gigaword, DUC2003, DUC2004).

OpenNMT framework (PyTorch).
- Replace 'softmax' with **sparsemax**, **Sparsegen-lin** and **Sparsehourglass**.
- Also varied temperature in softmax as another baseline.



SOURCE: *Roger Federer wins a record equaling sixth men's singles title at Aus Open on Sunday.*



λ control

Summary of Results:

Attention	TRANSLATION		SUMMARIZATION								
	FR-EN	EN-FR	Gigaword			DUC 2003			DUC 2004		
	BLEU	BLEU	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
softmax	36.38	36.00	34.80	16.64	32.15	27.95	**9.22**	24.54	30.68	12.24	28.12
softmax (with temp.)	36.63	**36.08**	35.00	17.15	32.57	27.78	8.91	24.53	31.64	**12.89**	28.51
sparsemax	36.73	35.78	34.89	16.88	32.20	27.29	8.48	24.04	30.80	12.01	28.04
sparsegen-lin	**37.27**	35.78	**35.90**	**17.57**	**33.37**	**28.13**	9.00	**24.89**	**31.85**	12.28	**29.13**
sparsehg	36.63	35.69	35.14	16.91	32.66	27.39	9.11	24.53	30.64	12.05	28.18

Key Contributions/Takeaways

- A unified framework for **sparse** probability mapping functions.
- Formulations *sparsegen-lin* and *sparsehourglass* – **control over sparsity**.
- **Convex hinge-based loss** functions for multilabel classification.
- **Sparser and more accurate prediction** for multilabel classification.
- Sparsity **control over attention heatmaps** in neural machine translation and abstractive summarization.

References

[1] André F. T. Martins and Ramón F. Astudillo. (2016) ***From softmax to sparsemax: A sparse model of attention and multi-label classification***. [ICML 2016]

[2] Alexandre de Brébisson and Pascal Vincent. (2016) ***An exploration of softmax alternatives belonging to the spherical loss family***. [ICLR 2016]

[3] Vlad Niculae and Mathieu Blondel. (2017) ***A regularized framework for sparse and structured neural attention***. [NIPS 2017]