Sparse Sequence-to-Sequence Models

Ben Peters — Instituto de Telecomunicações

→ **Vlad Niculae** — IT

André Martins — IT & Unbabel

github.com/deep-spin/entmax https://vene.ro @vnfrombucharest

(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention

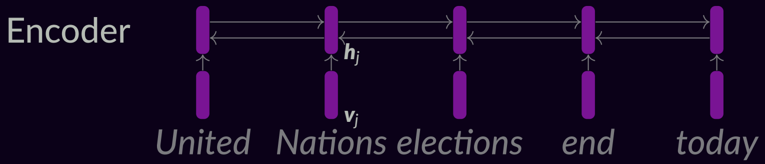
United Nations elections end today

Sequence-to-Sequence With Attention

Encoder

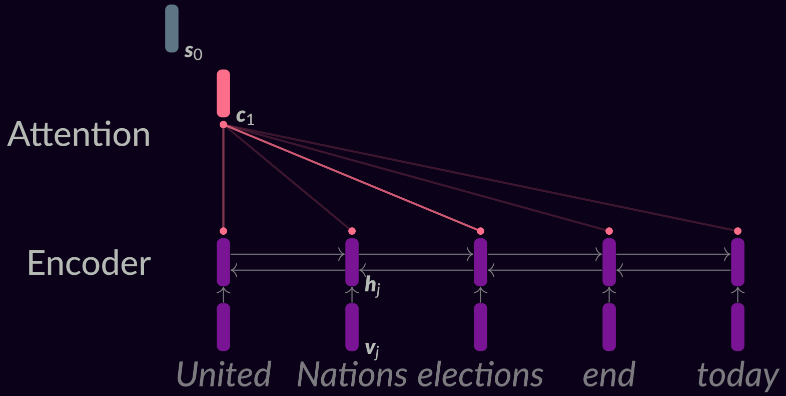
(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



attention weights computed with *softmax*:

for some decoder state $\boldsymbol{s}_t$, compute contextually weighted average of input $\boldsymbol{c}_t$:

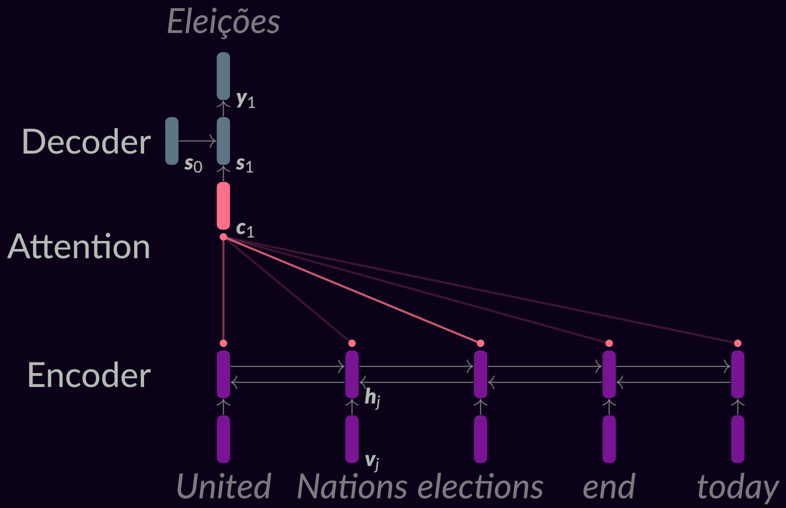
$$z_j = \boldsymbol{s}_t^\top \boldsymbol{W}^{(a)} \boldsymbol{h}_j$$

$$\pi_j = \mathrm{softmax}_j(\boldsymbol{z})$$

$$\boldsymbol{c}_t = \sum_j \pi_j \boldsymbol{h}_j$$

(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



predictive probability
(also using *softmax*!)

$$u_t = \tanh(\boldsymbol{W}^{(u)}[\boldsymbol{s}_t; \boldsymbol{c}_t])$$

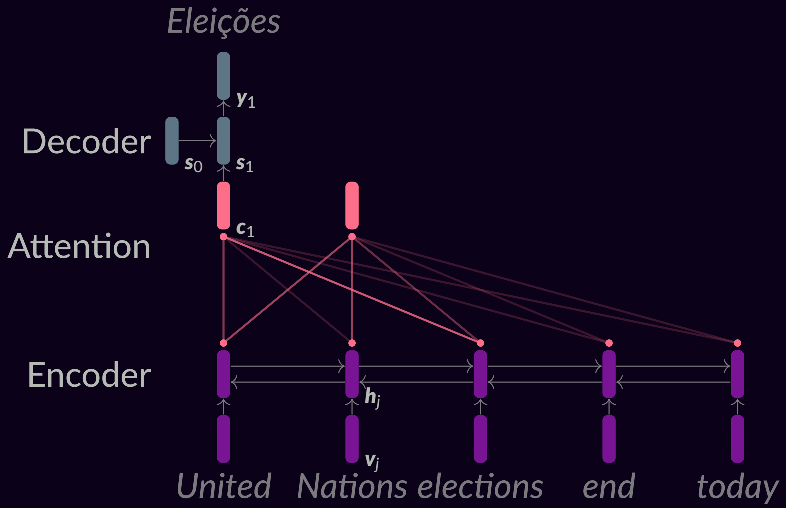
$$P(y_t \mid y_{1:t-1}, x) = \mathrm{softmax}(\boldsymbol{V}\boldsymbol{u}_t)$$

$P(y_1 \mid x)$

.70	Eleições
.11	Os
.10	As
.09	Nações
	...
10^{-6}	Bucarest

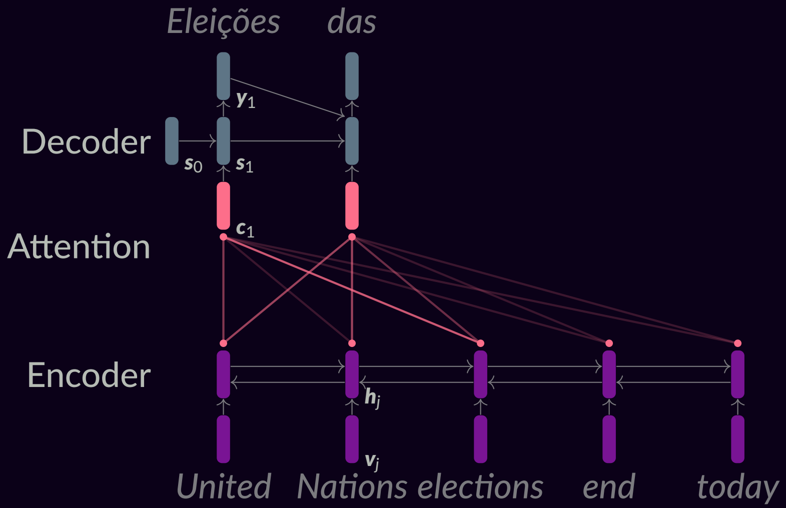
(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



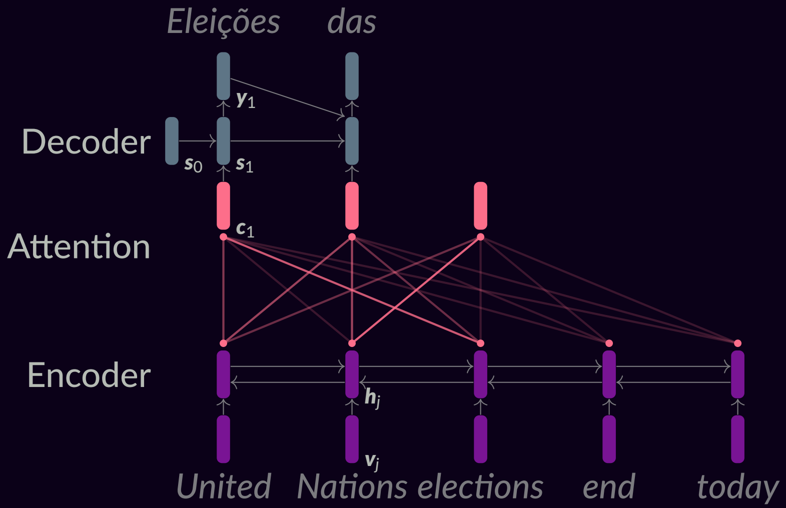
predictive probability

$P(y_2 \mid y_1, x)$

.40	das
.30	para
.20	ás
	...
10^{-7}	esquerda

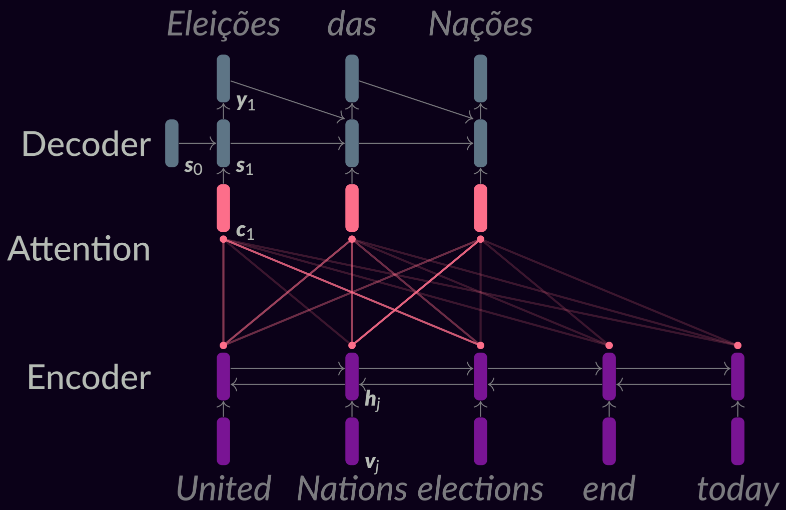
(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



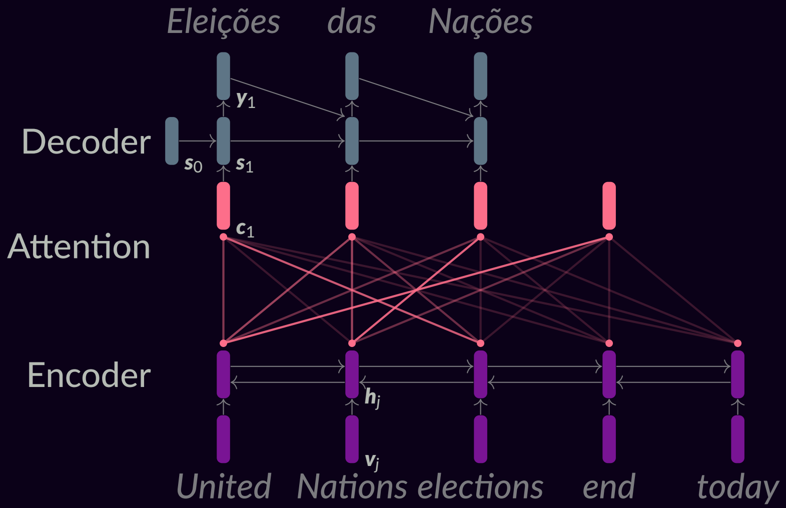
predictive probability

$P(y_3 \mid y_2, y_1, x)$

.80	Nações
.11	Representações
.03	assembleias
	...
10^{-8}	resultados

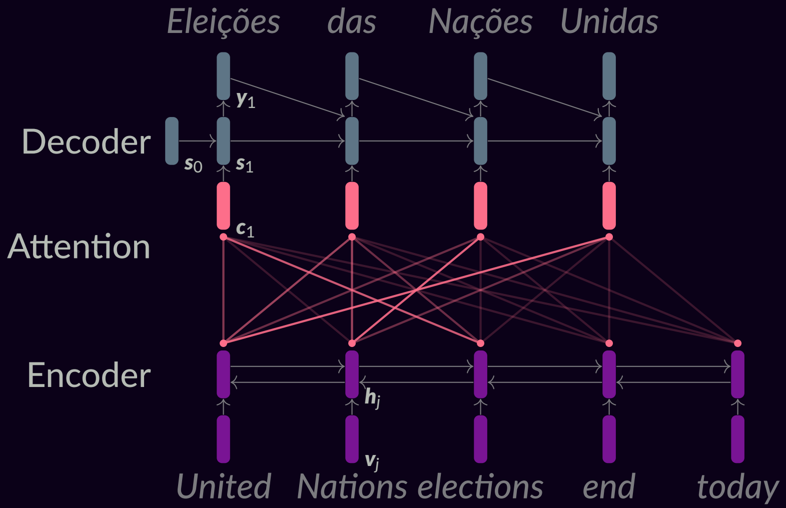
(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



(Bahdanau et al., 2015)

Sequence-to-Sequence With Attention



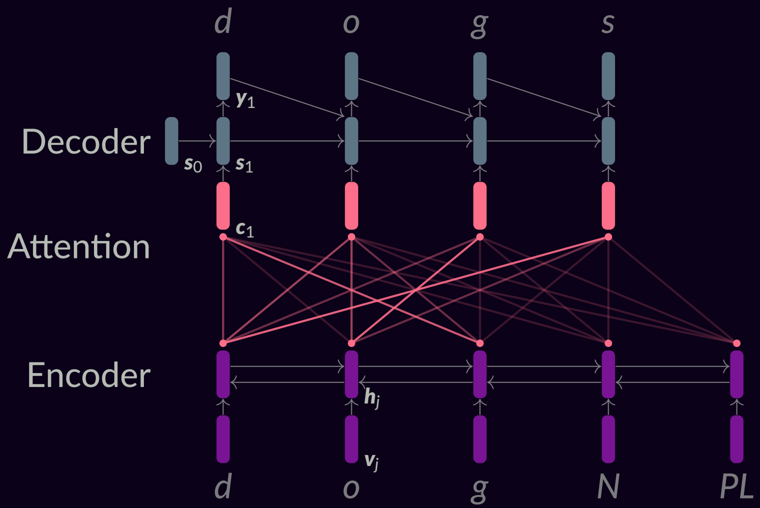
predictive probability

$P(y_4 \mid y_3, y_2, y_1, x)$

.90	Unidas
.05	Shopping
.01	,
	...
10^{-5}	aquático

(Bahdanau et al., 2015)

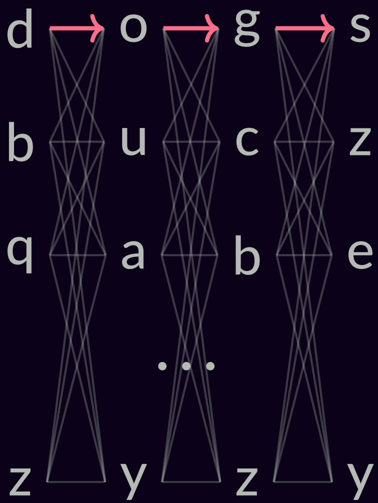
Sequence-to-Sequence With Attention



morphological inflection!

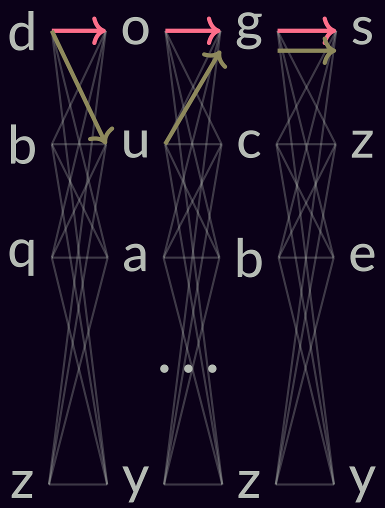
The Space of Outputs

The Space of Outputs



$p(\cdot) = 0.60$

The Space of Outputs



$p(\cdot) = 0.60$

$p(\cdot) = 0.13$

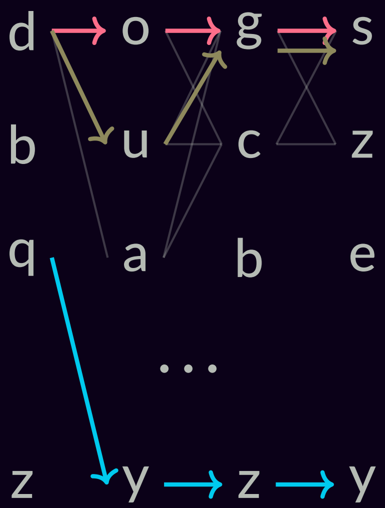
The Space of Outputs

$p(\cdot) = 0.60$

$p(\cdot) = 0.13$

$p(\cdot) = 10^{-4}$

The Space of Outputs: Made Sparse!



$p(\cdot) = 0.70$

$p(\cdot) = 0.20$

$p(\cdot) = 0$!!

Softmax plays two roles in seq2seq:

attention weights

for some decoder state $\boldsymbol{s}_t$, compute contextually weighted average of input $\boldsymbol{c}_t$:

$$z_j = \boldsymbol{s}_t^\top \boldsymbol{W}^{(a)} \boldsymbol{h}_j$$

$$\pi_j = \text{softmax}_j(\boldsymbol{z})$$

$$\boldsymbol{c}_t = \sum_j \pi_j \boldsymbol{h}_j$$

output probabilities

predict the probability of the next word:

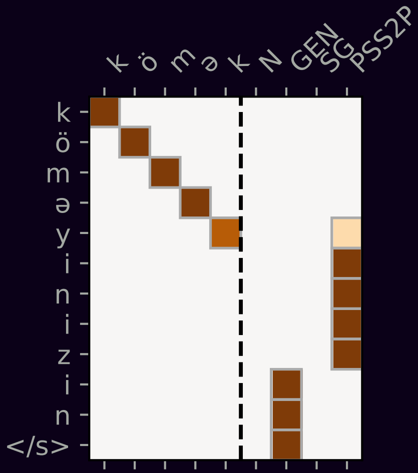
$$\boldsymbol{u}_t = \tanh(\boldsymbol{W}^{(u)}[\boldsymbol{s}_t; \boldsymbol{c}_t])$$

$$P(y_t \mid y_{1:t-1}, x) = \text{softmax}(\boldsymbol{V}\boldsymbol{u}_t)$$

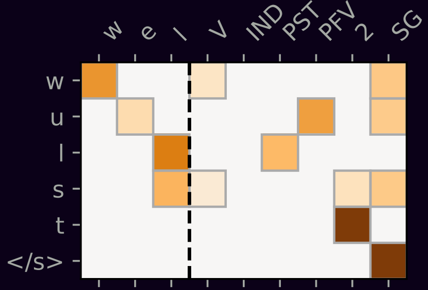
Our work: replace softmax with a *family of* new sparsity-inducing alternatives

(Peters and Martins, 2019)

Sparse Attention Weights / Alignments

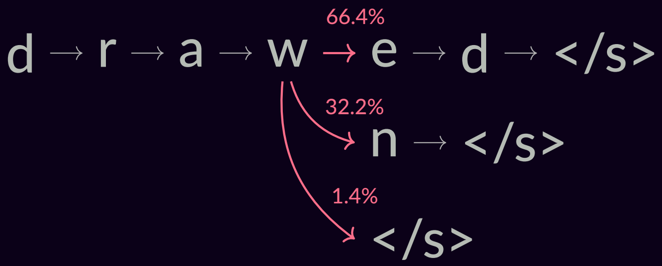


(Azeri)



(North Frisian)

Sparse Predictive Probabilities



What is softmax?

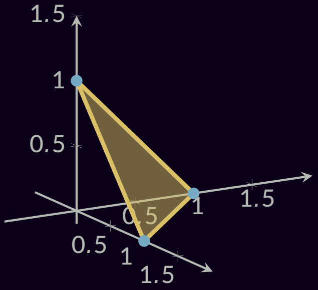
Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

What is softmax?

Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \ : \ \boldsymbol{p} \geq \boldsymbol{0}, \ \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices

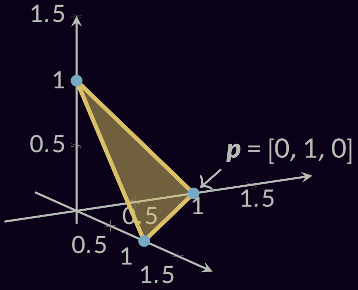


What is softmax?

Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \ : \ \boldsymbol{p} \geq \boldsymbol{0}, \ \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices

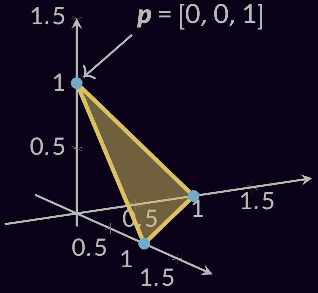


What is softmax?

Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \; : \; \boldsymbol{p} \geq \mathbf{0}, \; \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices

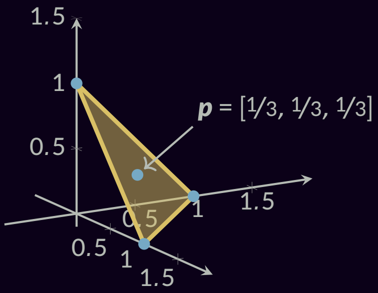


What is softmax?

Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \; : \; \boldsymbol{p} \geq \mathbf{0}, \; \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices



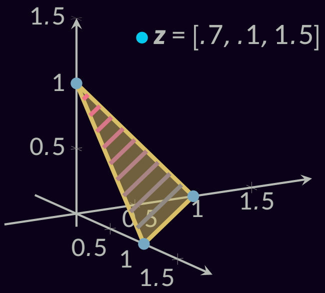
What is softmax?

Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \; : \; \boldsymbol{p} \geq \boldsymbol{0}, \; \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices

Expected score under $\boldsymbol{p}$: $\mathbb{E}_{i \sim \boldsymbol{p}}\, z_i = \boldsymbol{p}^\top \boldsymbol{z}$



What is softmax?

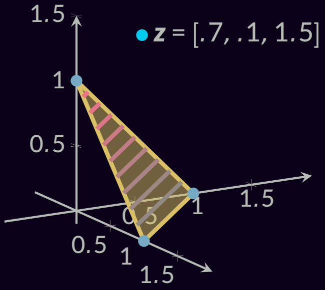
Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \ : \ \boldsymbol{p} \geq \boldsymbol{0}, \ \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices

Expected score under $\boldsymbol{p}$: $\mathbb{E}_{i \sim \boldsymbol{p}}\, z_i = \boldsymbol{p}^\top \boldsymbol{z}$

argmax



What is softmax?

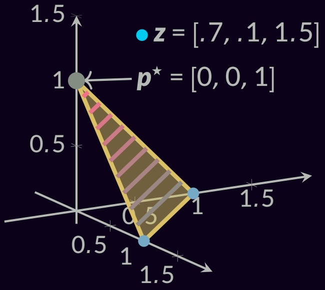
Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \ : \ \boldsymbol{p} \geq \boldsymbol{0}, \ \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices

Expected score under $\boldsymbol{p}$: $\mathbb{E}_{i \sim \boldsymbol{p}}\, z_i = \boldsymbol{p}^\top \boldsymbol{z}$

argmax maximizes **expected score**



$\underset{\boldsymbol{p} \in \triangle}{\arg\max}\, \boldsymbol{p}^\top \boldsymbol{z}$

What is softmax?

Often defined via $p_i := \frac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

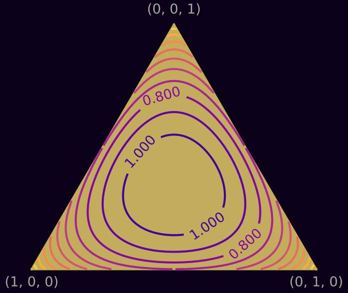
$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \ : \ \boldsymbol{p} \geq \mathbf{0}, \ \sum_j p_j = 1\}$

$\boldsymbol{p} \in \triangle$: probability distribution over choices

Expected score under $\boldsymbol{p}$: $\mathbb{E}_{i \sim \boldsymbol{p}} \, z_i = \boldsymbol{p}^\top \boldsymbol{z}$

argmax maximizes **expected score**

Shannon entropy of $\boldsymbol{p}$: $\mathrm{H}^{\mathrm{S}}(\boldsymbol{p}) := -\sum_i p_i \log p_i$



What is softmax?

Often defined via $p_i := \dfrac{\exp z_i}{\sum_j \exp z_j}$, but where does it come from?

$\triangle := \{\boldsymbol{p} \in \mathbb{R}^d \;:\; \boldsymbol{p} \geq \boldsymbol{0},\; \sum_j p_j = 1\}$

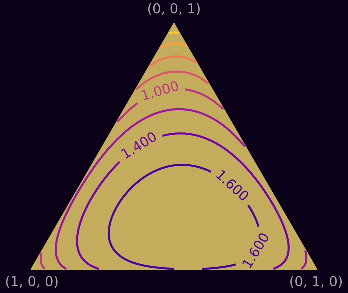
$\boldsymbol{p} \in \triangle$: probability distribution over choices

Expected score under $\boldsymbol{p}$: $\mathbb{E}_{i \sim \boldsymbol{p}}\, z_i = \boldsymbol{p}^\top \boldsymbol{z}$

argmax maximizes **expected score**

Shannon entropy of $\boldsymbol{p}$: $\mathrm{H}^{\mathrm{s}}(\boldsymbol{p}) := -\sum_i p_i \log p_i$

softmax maximizes **expected score** + **entropy**:

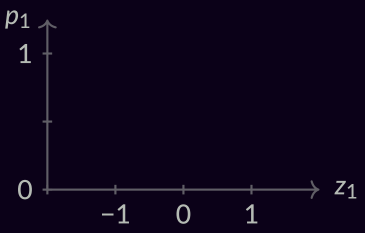


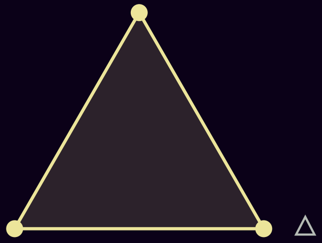
$$\arg\max_{\boldsymbol{p} \in \triangle} \boldsymbol{p}^\top \boldsymbol{z} + \mathrm{H}^{\mathrm{s}}(\boldsymbol{p})$$

(Niculae and Blondel, 2017; Blondel et al., 2019)

Generalizing Softmax Using Entropies

$$\boldsymbol{\pi}_{\mathrm{H}}(\boldsymbol{z}) = \arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^{\top}\boldsymbol{z} + \mathrm{H}(\boldsymbol{p})$$

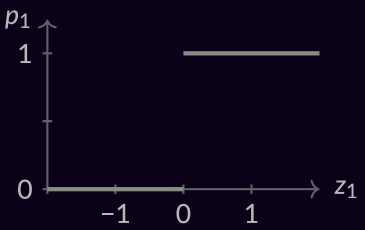


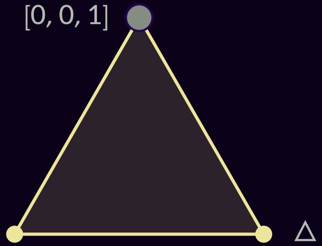


Generalizing Softmax Using Entropies

$$\boldsymbol{\pi}_{\mathrm{H}}(\boldsymbol{z}) = \arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top \boldsymbol{z} + \mathrm{H}(\boldsymbol{p})$$

- argmax: $\mathrm{H}^0(\boldsymbol{p}) = 0$



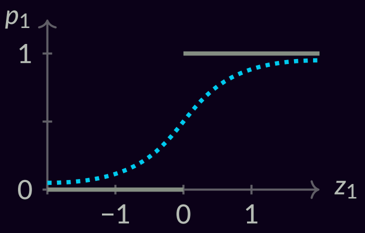


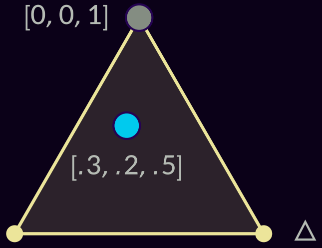
(Niculae and Blondel, 2017; Blondel et al., 2019)

Generalizing Softmax Using Entropies

$$\boldsymbol{\pi}_{\mathrm{H}}(\boldsymbol{z}) = \underset{\boldsymbol{p}\in\triangle}{\arg\max}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathrm{H}(\boldsymbol{p})$$

- argmax: $\mathrm{H}^0(\boldsymbol{p}) = 0$
- softmax: $\mathrm{H}^s(\boldsymbol{p}) = -\sum_j p_j \log p_j$

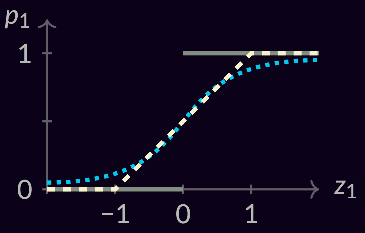


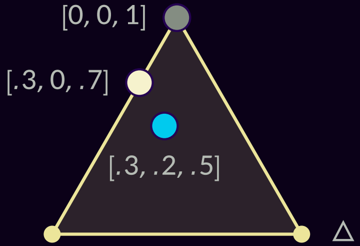


Generalizing Softmax Using Entropies

$$\pi_H(z) = \arg\max_{p \in \triangle} p^\top z + H(p)$$

- argmax: $H^0(p) = 0$
- softmax: $H^s(p) = -\sum_j p_j \log p_j$
- sparsemax: $H^g(p) = 1/2 \sum_j p_j(1 - p_j)$





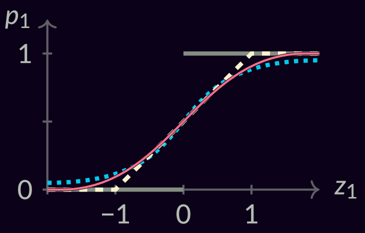
(Martins and Astudillo, 2016)

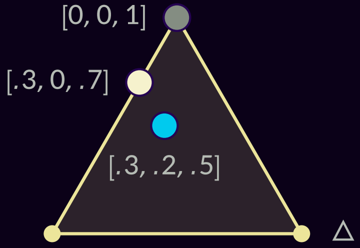
(Niculae and Blondel, 2017; Blondel et al., 2019)

Generalizing Softmax Using Entropies

$$\boldsymbol{\pi}_{\mathrm{H}}(\boldsymbol{z}) = \underset{\boldsymbol{p}\in\triangle}{\arg\max}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathrm{H}(\boldsymbol{p})$$

- argmax: $\mathrm{H}^0(\boldsymbol{p}) = 0$
- softmax: $\mathrm{H}^{\mathrm{s}}(\boldsymbol{p}) = -\sum_j p_j \log p_j$
- sparsemax: $\mathrm{H}^{\mathrm{g}}(\boldsymbol{p}) = 1/2 \sum_j p_j(1-p_j)$
- α-entmax: $\mathrm{H}^{\mathrm{t}}_{\alpha}(\boldsymbol{p}) = \frac{1}{\alpha(\alpha-1)} \sum_j (p_j - p_j^{\alpha})$





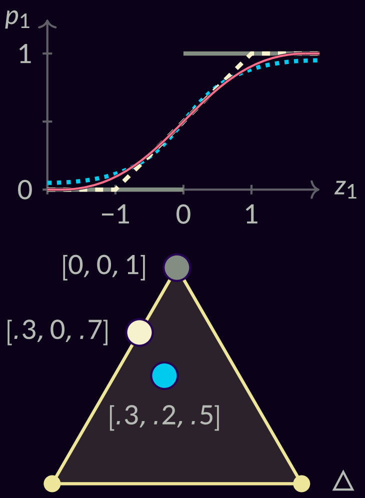
(Niculae and Blondel, 2017; Blondel et al., 2019)

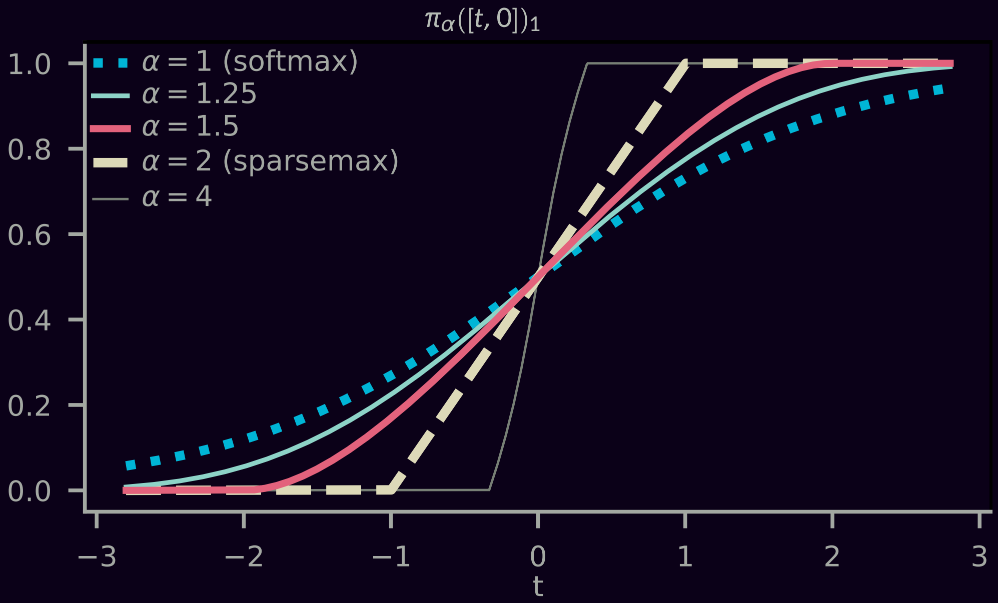
Generalizing Softmax Using Entropies

$$\boldsymbol{\pi}_{\mathrm{H}}(\boldsymbol{z}) = \underset{\boldsymbol{p}\in\triangle}{\arg\max}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathrm{H}(\boldsymbol{p})$$

- argmax: $\mathrm{H}^0(\boldsymbol{p}) = 0$
- softmax: $\mathrm{H}^{\mathrm{s}}(\boldsymbol{p}) = -\sum_j p_j \log p_j$
- sparsemax: $\mathrm{H}^{\mathrm{g}}(\boldsymbol{p}) = 1/2 \sum_j p_j(1 - p_j)$
- α-entmax: $\mathrm{H}^{\mathrm{t}}_\alpha(\boldsymbol{p}) = \frac{1}{\alpha(\alpha-1)} \sum_j (p_j - p_j^\alpha)$

Tsallis α-entropy (Tsallis, 1988).
Depicted: $\alpha = 1.5$. Uncovers softmax ($\alpha \to 1$) and sparsemax ($\alpha = 2$).




$\pi_\alpha([t, 0])_1$
$\alpha = 1$ (softmax)
$\alpha = 1.25$
$\alpha = 1.5$
$\alpha = 2$ (sparsemax)
$\alpha = 4$
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
t

Computing α-entmax

$$\boldsymbol{\pi}_{\mathsf{H}^{\mathsf{t}}_{\alpha}}(\boldsymbol{z}) := \underset{\boldsymbol{p}\in\triangle}{\arg\max}\, \boldsymbol{p}^\top \boldsymbol{z} + \mathsf{H}^{\mathsf{t}}_{\alpha}(\boldsymbol{p})$$

Solution has the form:

$$\boldsymbol{\pi}_{\mathsf{H}^{\mathsf{t}}_{\alpha}}(\boldsymbol{z}) = \left[(\alpha - 1)\boldsymbol{z} - \tau\mathbf{1}\right]_+^{1/\alpha-1}$$

Algorithms:

bisection

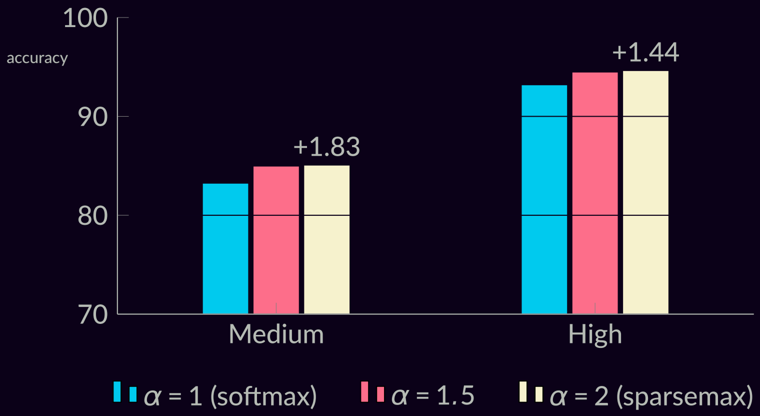
- approximate; bracket $\tau \in [\tau_{\text{lo}}, \tau_{\text{hi}}]$
- gain 1 bit per $\mathcal{O}(d)$ iteration
- `float32` has 23 mantissa bits

sort-based

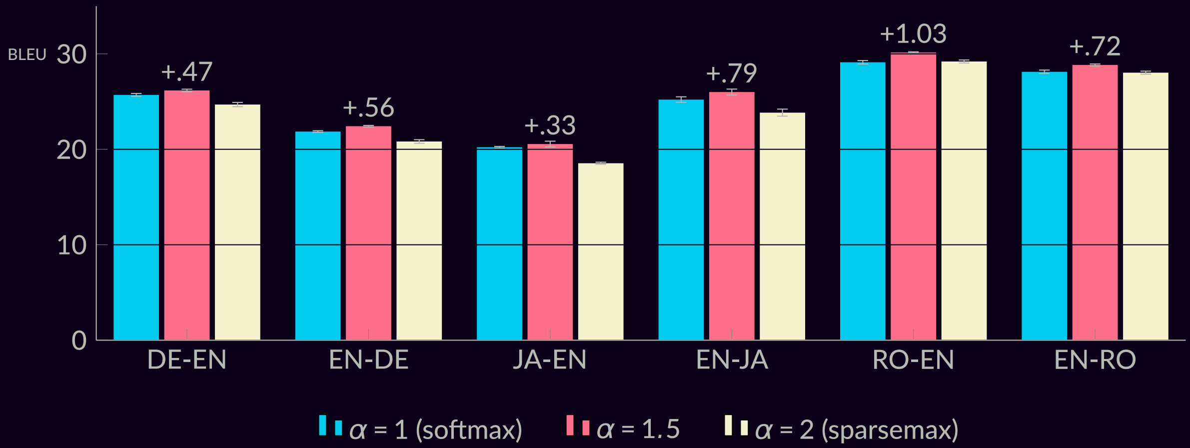
- exact algorithm, $\mathcal{O}(d \log d)$
- available only for $\alpha \in \{1.5, 2\}$
- huge speed-up from partial sorting when expecting sparse solutions

Morphological inflection

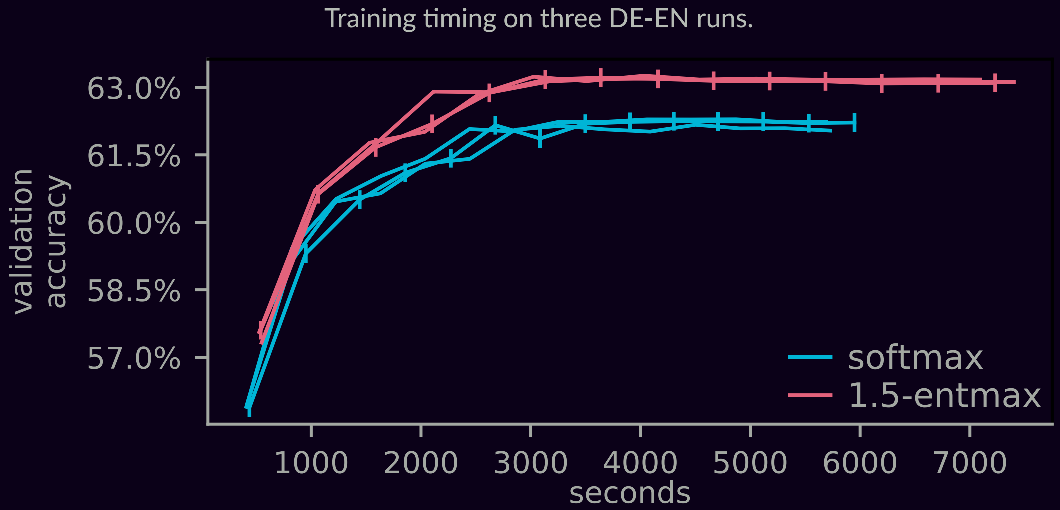
SIGMORPHON 2018. Shared multi-lingual model.
Medium: 1k pairs per language, 102 languages.
High: 10k pairs per language, 86 languages.



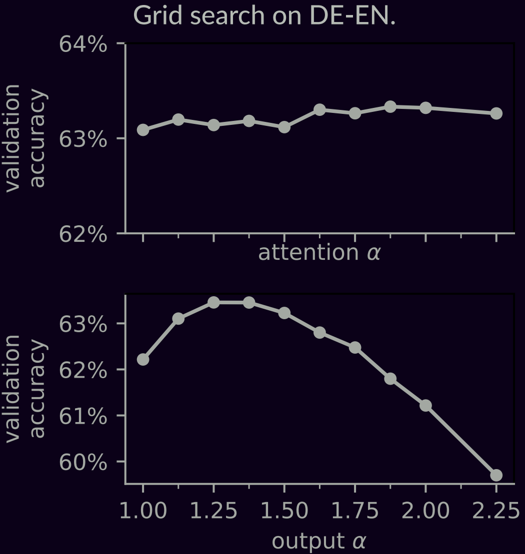
Neural Machine Translation



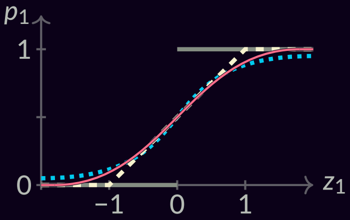
Sparse Mappings Don't Slow Down Training



Impact of Fine Tuning α

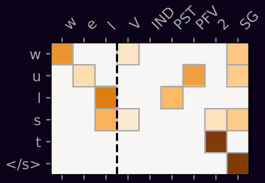


Sparse Seq2Seq: Conclusions

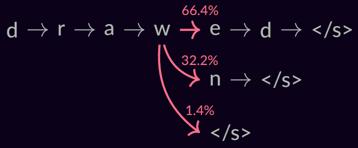


New family of sparse mappings α-entmax,
algorithms for efficient forward & backward passes.

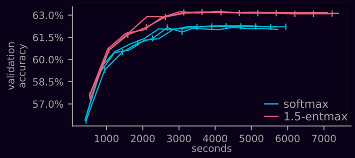
sparse attention weights



sparse output space



performance improvements



vlad@vene.ro github.com/deep-spin/entmax https://vene.ro @vnfrombucharest

Acknowledgements

This work was supported by the European Research Council (ERC StG DeepSPIN 758969) and by the Fundação para a Ciência e Tecnologia through contract UID/EEA/50008/2019 and CMUPERI/TIC/0046/2014 (GoLocal).

Some icons by Dave Gandy and Freepik via flaticon.com.

References I

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). “Neural machine translation by jointly learning to align and translate”. In: *Proc. of ICLR*.

Blondel, Mathieu, André FT Martins, and Vlad Niculae (2019). “Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms”. In: *Proc. AISTATS*.

Martins, André FT and Ramón Fernandez Astudillo (2016). “From softmax to sparsemax: A sparse model of attention and multi-label classification”. In: *Proc. of ICML*.

Niculae, Vlad and Mathieu Blondel (2017). “A regularized framework for sparse and structured neural attention”. In: *Proc. of NeurIPS*.

Peters, Ben and André FT Martins (2019). “IT-IST at the SIGMORPHON 2019 shared task: Sparse two-headed models for inflection”. In: *Proc. SIGMORPHON*.

Tsallis, Constantino (1988). “Possible generalization of Boltzmann-Gibbs statistics”. In: *Journal of Statistical Physics* 52, pp. 479–487.