From Sparse Modeling to Sparse Communication

André Martins

Mathematics, Physics, and Machine Learning Seminar, February 24, 2022

Our Amazing Team (December 2019, pre-COVID)

DeepSPIN

- ERC starting grant (2018–23)
- Goal: put together *deep learning* and *structured prediction* for *natural language processing*
- More details: https://deep-spin.github.io

From Sparse Modeling ...

- Mostly used with linear models, lots of work in the 2000s
- Main idea: embed a sparse regularizer (e.g. ℓ_1-norm) in the learning objective
- Irrelevant features get zero weight and can be discarded
- Extensions to structured sparsity (group-lasso, fused-lasso, etc.)

... to Sparse Communication:

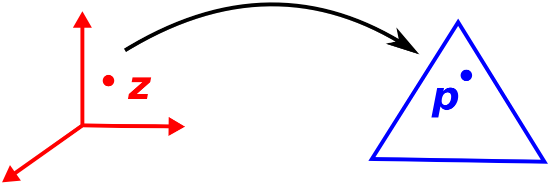
- Mostly used with neural networks, most work after 2015
- Main idea: sparse neuron activations (biological plausibility)
- Predictions are triggered by a few neurons only (input-dependent)
- Example: ReLUs, dropout, sparse attention mechanisms

This Talk

An inventory of transformations that capture **sparsity** and **structure**:

- All differentiable (efficient forward and backward propagation)
- Can be used at hidden (attention) or output layers (loss)
- Can make a bridge between the continuous and discrete worlds
- Effective in several natural language processing tasks.

Building block:



Sparse transformations from the Euclidean space to the simplex $\triangle$.

$\sum$ vs. $\int$

Commonly we have to opt between discrete or continuous models:

- Language is symbolic and *discrete*
- Neural networks use (and learn) *continuous* representations

We should look at what happens in-between!

Sparsity might help with this, but...

$\sum$ vs. $\int$

Commonly we have to opt between discrete or continuous models:

- Language is symbolic and *discrete*
- Neural networks use (and learn) *continuous* representations

We should look at what happens in-between!

Sparsity might help with this, but...

... sparse probabilities are understudied and often excluded from theory:

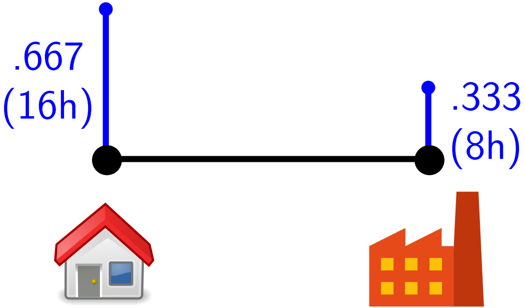
- Hammersley-Clifford theorem in graphical models
- Pitman-Koopman-Darmois theorem (sufficient statistics and exponential families)
- Log-likelihood is $-\infty$ if estimated probability is 0.

Motivating Example: John's Life

John splits his day as follows: he works 8h/day, and stays home 16h/day.

Motivating Example: John's Life

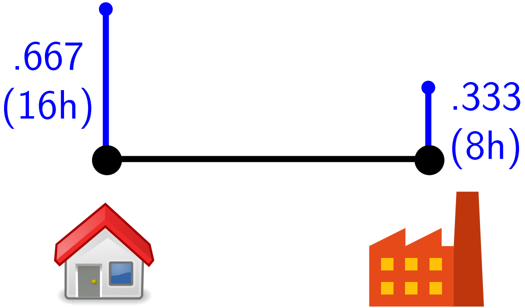
John splits his day as follows: he works 8h/day, and stays home 16h/day.



Motivating Example: John's Life

John splits his day as follows: he works 8h/day, and stays home 15h/day.

He is in transit 1h/day to commute to work and back.



Motivating Example: John's Life

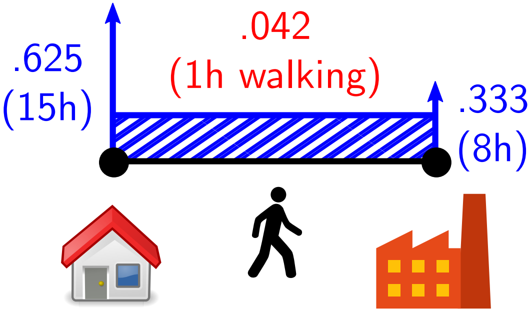
John splits his day as follows: he works 8h/day, and stays home 15h/day.

He is in transit 1h/day to commute to work and back.

Motivating Example: John's Life

John splits his day as follows: he works 8h/day, and stays home 15h/day.

He is in transit 1h/day to commute to work and back.



Is John's location a *discrete* or *continuous* random variable?

Motivating Example: John's Life

John splits his day as follows: he works 8h/day, and stays home 15h/day.

He is in transit 1h/day to commute to work and back.

Is John's location a *discrete* or *continuous* random variable? It's mixed.

Outline

Recap: Softmax and Argmax

Softmax exponentiates and normalizes:

$$\mathrm{softmax}(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z})}{\sum_{k=1}^{K} \exp(z_k)}$$

- **Fully dense:** $\mathrm{softmax}(\boldsymbol{z}) > 0, \forall \boldsymbol{z}$
- Used both as a loss function (cross-entropy) and for attention.

Recap: Softmax and Argmax

Softmax exponentiates and normalizes:

$$\mathrm{softmax}(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z})}{\sum_{k=1}^{K} \exp(z_k)}$$

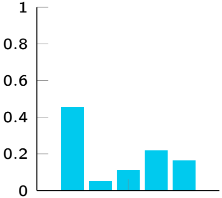
- **Fully dense:** $\mathrm{softmax}(\boldsymbol{z}) > 0, \forall \boldsymbol{z}$
- Used both as a loss function (cross-entropy) and for attention.

Argmax can be written as:

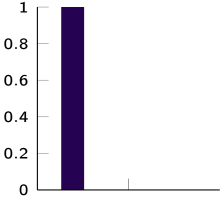
$$\begin{aligned} \mathrm{argmax}(\boldsymbol{z}) &:= \arg\max_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} \\ &= \lim_{\tau \to 0^+} \mathrm{softmax}(\boldsymbol{z}/\tau) \qquad \text{(temperature trick)} \end{aligned}$$

- Retrieves a **one-hot vector** for the highest scored index.

softmax($\boldsymbol{z}$)

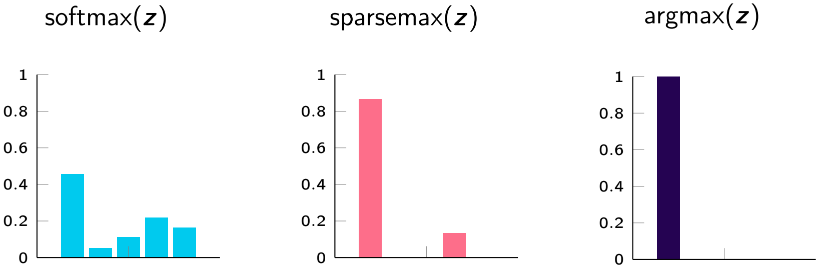


argmax($\boldsymbol{z}$)



(Same $\boldsymbol{z} = [1.0716, -1.1221, -0.3288, 0.3368, 0.0425]$)

- Argmax is an extreme case of sparsity, but it is **discontinuous**.
- Is there a **sparse** and **differentiable** alternative?



(Same $z = [1.0716, -1.1221, -0.3288, 0.3368, 0.0425]$)

- Argmax is an extreme case of sparsity, but it is **discontinuous**.
- Is there a **sparse** and **differentiable** alternative?

Sparsemax (Martins and Astudillo, 2016, ICML)

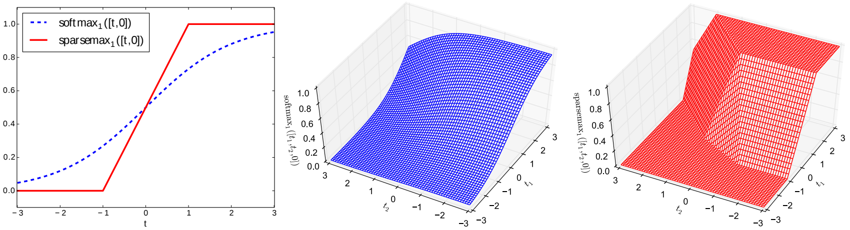
Euclidean projection of $\boldsymbol{z}$ onto the probability simplex $\triangle$:

$$\begin{aligned} \mathrm{sparsemax}(\boldsymbol{z}) &:= \arg\min_{\boldsymbol{p}\in\triangle} \|\boldsymbol{p}-\boldsymbol{z}\|^2 \\ &= \arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{z}^\top\boldsymbol{p} - \frac{1}{2}\|\boldsymbol{p}\|^2. \end{aligned}$$

- Likely to hit the boundary of the simplex, in which case $\mathrm{sparsemax}(\boldsymbol{z})$ becomes sparse (hence the name)
- End-to-end differentiable
- Forward pass: $O(K \log K)$ or $O(K)$, **(almost) as fast as softmax**
- Backprop: sublinear, **better than softmax!**

Sparsemax in 2D and 3D

(Martins and Astudillo, 2016, ICML)



- Sparsemax is piecewise linear, but asymptotically similar to softmax.

Ω-Regularized Argmax (Niculae and Blondel, 2017, NeurIPS)

For convex Ω, define the **Ω-regularized argmax transformation**:

$$\mathrm{argmax}_{\Omega}(\boldsymbol{z}) := \arg\max_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$$

- **Argmax** corresponds to **no regularization**, $\Omega \equiv 0$
- **Softmax** amounts to **entropic regularization**, $\Omega(\boldsymbol{p}) = \sum_{i=1}^{K} p_i \log p_i$
- **Sparsemax** amounts to **ℓ_2-regularization**, $\Omega(\boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2$

Is there something in-between?

Entmax (Peters et al., 2019, ACL)

Parametrized by $\alpha \geq 0$:

$$\Omega_\alpha(\boldsymbol{p}) := \begin{cases} \frac{1}{\alpha(\alpha-1)} \left(1 - \sum_{i=1}^{K} p_i^\alpha\right) & \text{if } \alpha \neq 1 \\ \sum_{i=1}^{K} p_i \log p_i & \text{if } \alpha = 1. \end{cases}$$

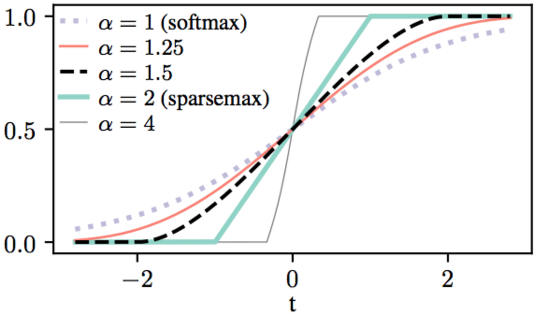
Related to **Tsallis generalized entropies** (Tsallis, 1988).

- **Argmax** corresponds to $\alpha \to \infty$
- **Softmax** amounts to $\alpha \to 1$
- **Sparsemax** amounts to $\alpha = 2$.

Key result: always sparse for $\alpha > 1$, sparsity increases with α

- Forward pass for general α can be done with a bissection algorithm
- Backward pass runs in sublinear time.

Entmax in 2D (Peters et al., 2019, ACL)



$\alpha = 1.5$ is a sweet spot!

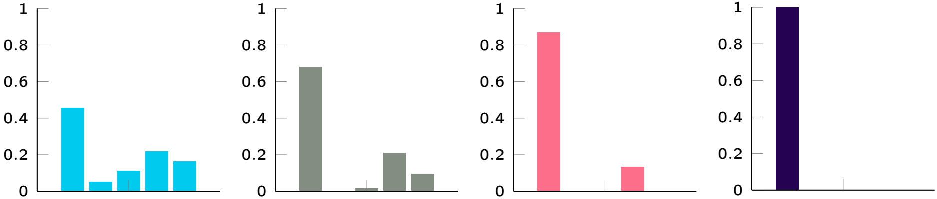
- Efficient exact algorithm (nearly as fast as softmax), smooth, and good empirical performance.

Pytorch code: `https://github.com/deep-spin/entmax`

Sparse Transformations (Peters et al., 2019, ACL)

$\alpha = 1$ | $\alpha = 1.5$ | $\alpha = 2$ | $\alpha = \infty$

$\mathrm{softmax}(\boldsymbol{z})$ | $1.5\text{-}\mathrm{entmax}(\boldsymbol{z})$ | $\mathrm{sparsemax}(\boldsymbol{z})$ | $\mathrm{argmax}(\boldsymbol{z})$

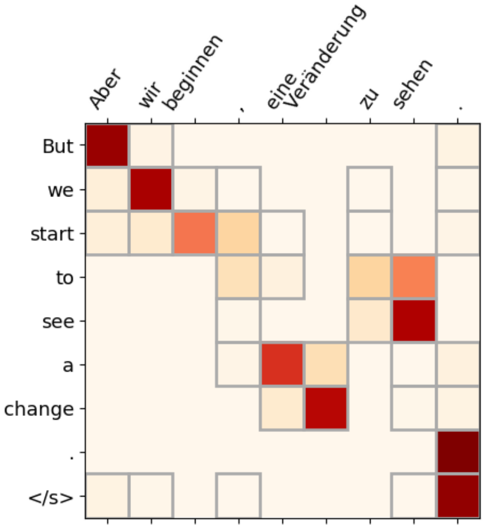


(Same $\boldsymbol{z} = [1.0716, -1.1221, -0.3288, 0.3368, 0.0425]$)

Example: Sparse Attention for Machine Translation

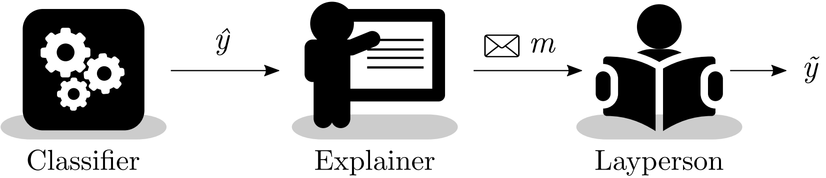
(Peters et al., 2019, ACL)

- **Selects** source words when generating a target word (sparse alignments)
- Better interpretability
- Can also model fertility: **constrained** sparsemax
  (Malaviya et al., 2018, ACL)
- Can also learn α (**adaptively sparse** transformers):
  (Correia et al., 2019, EMNLP)



Example: Sparse Attention for Explainability

(Treviso and Martins, 2020, BlackboxNLP)



- A classifier makes a prediction
- An "explainer" (embedded or not in the classifier) generates a sparse **message** that explains the classifier's decision
- The layperson receives the message and tries to guess the classifier's prediction (also called *simulatability*, *forward simulation/prediction*)
- Communication success rate: how often the two predictions match?

From Sparse Modeling to Sparse Communication

(Treviso and Martins, 2020, BlackboxNLP)

	Model interpretability	Prediction explainability
Wrappers	• Forward selection • Backward elimination (Kohavi and John, 1997)	• Input reduction (Feng et al., 2018) • Erasure (leave-one-out) (Li et al., 2016b; Serrano and Smith, 2019) • LIME (Ribeiro et al., 2016)
Filters	• PMI (Church and Hanks, 1990) • recursive feature elimination (Guyon et al., 2002)	• Input gradient (Li et al., 2016a) • LRP (Bach et al., 2015) • top-k softmax attention
Embedded	• ℓ_1-regularization (Tibshirani, 1996) • elastic net (Zou and Hastie, 2005)	• Stochastic attention (Xu et al., 2015; Lei et al., 2016; Bastings et al., 2019) • Sparse attention

Other Related Transformations

Constrained softmax **(Martins and Kreutzer, 2017, EMNLP)**,

Constrained sparsemax **(Malaviya et al., 2018, ACL)**:

- Allows placing a **budget** on how much attention a word can receive
- Useful to model **fertility** in machine translation

Fusedmax **(Niculae and Blondel, 2017, NeurIPS)**:

- Can promote **structured sparsity** (contiguous selection)

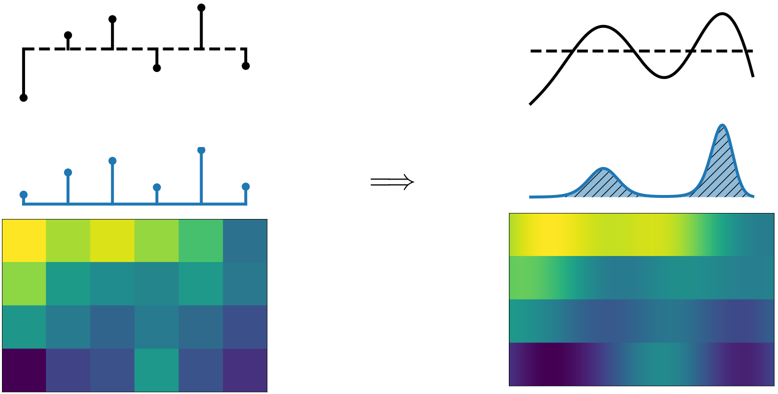
(LP-)SparseMAP **Niculae et al. (2018, ICML)**, **Niculae and Martins (2020, ICML)**:

- Extends sparsemax to **sparse structured prediction**.
- Can be used as hidden differentiable layer or output layer.
- Works with arbitrary factor graph (e.g. logic constraints).

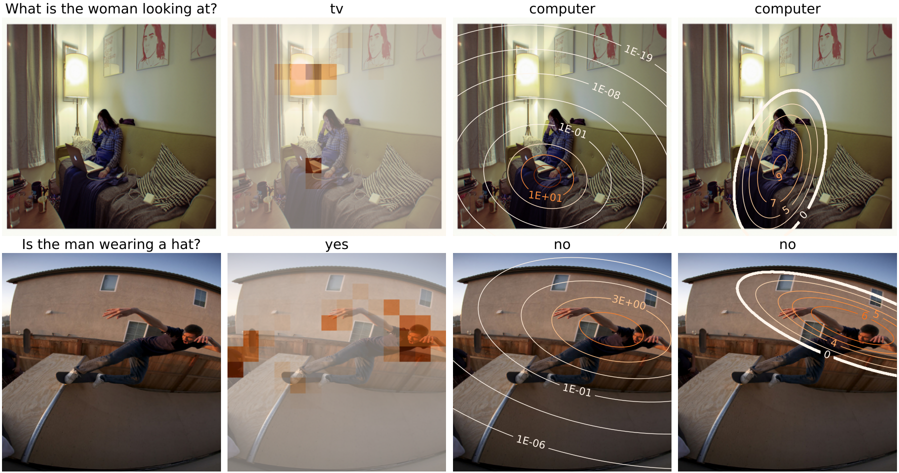
Sparse and Continuous Attention

(Martins et al., 2020a, NeurIPS)

- So far: attention over a **finite set** (words, pixel regions, etc.)
- We generalize attention to *arbitrary sets*, possibly continuous.



Example: Visual Question Answering



(original image) (discrete attention) (continuous softmax) (continuous sparsemax)

Outline

Entmax Losses

- Entmax can also be used as a loss in the **output layer** (to replace logistic/cross-entropy loss)
- However, not expressed as a log-likelihood (which could lead to $\log(0)$ problems due to sparsity)
- Instead, we build a entmax loss inspired by **Fenchel-Young losses**.

Softmax: Logistic Loss

- Softmax gives us the **logistic loss** (or negative log-likelihood):

$$L_{\text{softmax}}(\boldsymbol{z}; k) \;=\; -\log \operatorname{softmax}_k(\boldsymbol{z}) \;=\; -z_k + \log \textstyle\sum_j \exp(z_j),$$

- Great! Can we do the same to define a loss for sparsemax?
- Unfortunately, log-likelihood does not work well with sparsemax: labels with *exactly* zero probability would make log-likelihood $-\infty$...

Sparsemax Loss

- A better approach: construct an alternative loss function whose *gradient* resembles the gradient of the logistic loss:

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

- Looks a bizarre idea, but we'll motivate it later!
- So, by design, we want $L_{\text{sparsemax}}$ to be differentiable and such that:

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

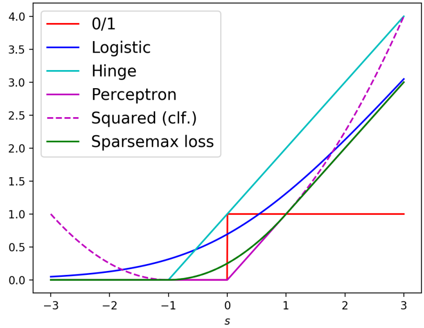
- This property is fulfilled by the following function (**sparsemax loss**):

$$L_{\text{sparsemax}}(\boldsymbol{z}; k) = -z_k + \boldsymbol{z}^\top \text{sparsemax}(\boldsymbol{z}) - \frac{1}{2}\|\text{sparsemax}(\boldsymbol{z})\|^2 + \frac{1}{2}.$$

Two Dimensions: Relation to the Huber Loss

- In 2D, $L_{\text{sparsemax}}$ reduces to the **Huber classification loss** known from robust statistics (Huber, 1964; Zhang, 2004)
- Let the correct label be $k = 1$, and define $s = z_2 - z_1$:

$$L_{\text{sparsemax}}(s) = \begin{cases} 0 & \text{if } s \leq -1 \\ \frac{(s-1)^2}{4} & \text{if } -1 < s < 1 \\ s & \text{if } s \geq 1 \end{cases}$$



Recap: Ω-Regularized Argmax (Niculae and Blondel, 2017, NeurIPS)

For convex Ω, define the Ω**-regularized argmax transformation**:

$$\mathrm{argmax}_{\Omega}(\boldsymbol{z}) := \arg\max_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$$

- **Argmax** corresponds to **no regularization**, $\Omega \equiv 0$
- **Softmax** amounts to **entropic regularization**, $\Omega(\boldsymbol{p}) = \sum_{i=1}^{K} p_i \log p_i$
- **Sparsemax** amounts to ℓ_2**-regularization**, $\Omega(\boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2$

All these are particular cases of α-entmax (Peters et al., 2019, ACL).

Fenchel-Young Losses (Blondel et al., 2020, JMLR)

Assess compatibility between groundtruth $\boldsymbol{q} \in \triangle$ and scores $\boldsymbol{z} \in \mathbb{R}^K$

Convex conjugate $\Omega^*(\boldsymbol{z}) := \max_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$

$$L_\Omega(\boldsymbol{z}, \boldsymbol{q}) := \Omega^*(\boldsymbol{z}) + \Omega(\boldsymbol{q}) - \boldsymbol{z}^\top \boldsymbol{q}$$

Fenchel-Young Losses (Blondel et al., 2020, JMLR)

Assess compatibility between groundtruth $\boldsymbol{q} \in \triangle$ and scores $\boldsymbol{z} \in \mathbb{R}^K$

Convex conjugate $\Omega^*(\boldsymbol{z}) := \max_{\boldsymbol{p} \in \triangle} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$

$$L_\Omega(\boldsymbol{z}, \boldsymbol{q}) := \Omega^*(\boldsymbol{z}) + \Omega(\boldsymbol{q}) - \boldsymbol{z}^\top \boldsymbol{q}$$

Properties:

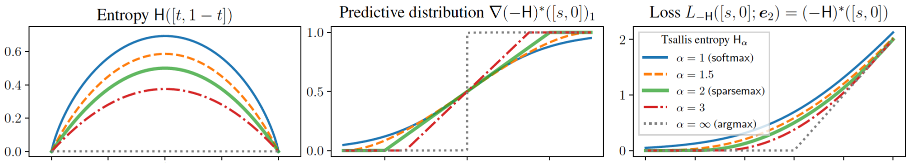
- $L_\Omega(\boldsymbol{z}, \boldsymbol{q}) \geq 0$ (automatic from **Fenchel-Young inequality**)
- $L_\Omega(\boldsymbol{z}, \boldsymbol{q}) = 0$ iff $\boldsymbol{q} = \operatorname{argmax}_\Omega(\boldsymbol{z})$
- L_Ω is convex and differentiable with $\nabla L_\Omega(\boldsymbol{z}, \boldsymbol{q}) = \operatorname{argmax}_\Omega(\boldsymbol{z}) - \boldsymbol{q}$

Recovers **cross-entropy loss**, **sparsemax loss**, and many other known losses

Also called "mixed-type Bregman divergences" (Amari, 2016).

Entmax Transformations and Losses

(Blondel et al., 2020, JMLR)

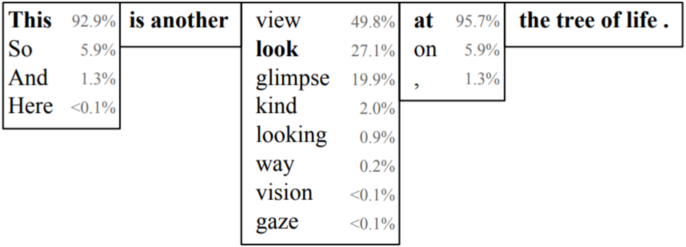


- Key result: for all $\alpha > 1$, all transformations are **sparse** and lead to losses with **margins**!
- The **margin size** is related to the **slope** of the entropy in the simplex corners! ($\frac{1}{\alpha - 1}$ for entmax losses.)
- See paper for details!

Pytorch code: `https://github.com/deep-spin/entmax`

Example: Machine Translation

(Peters et al., 2019, ACL) (Peters and Martins, 2021, NAACL)



(Source: "Dies ist ein weiterer Blick auf den Baum des Lebens.")

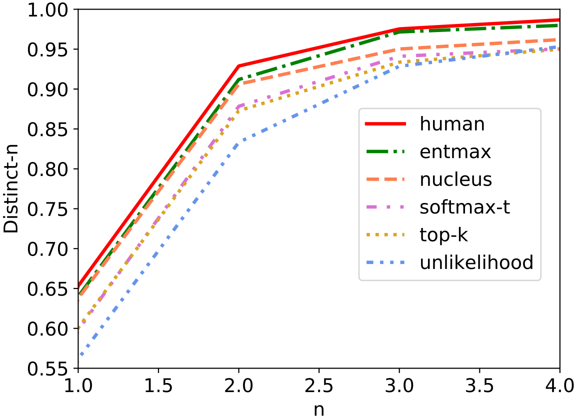
- Only a few words get non-zero probability at each time step
- Auto-completion when several words in a row have probability 1
- Useful for predictive translation.

Entmax Sampling (Martins et al., 2020b, EMNLP)

Use the entmax loss for training language models.

At test time, sample from this sparse distribution.

Better quality with less repetitions than other methods:



Outline

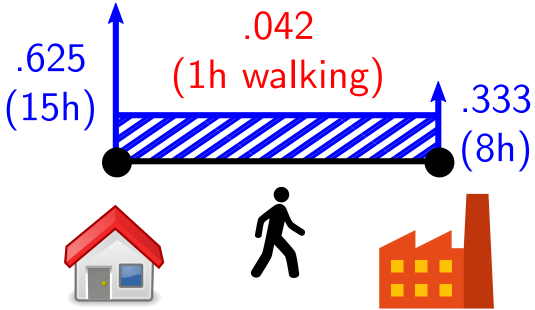
Mixed Distributions (Farinhas et al., 2022, ICLR)

- We saw how to obtain sparse probability distributions.
- How can we use them to bridge the gap between *discrete* and *continuous* domains?
- We'll see how next.

Back to John's Life

John splits his day as follows: he works 8h/day, and stays home 15h/day.

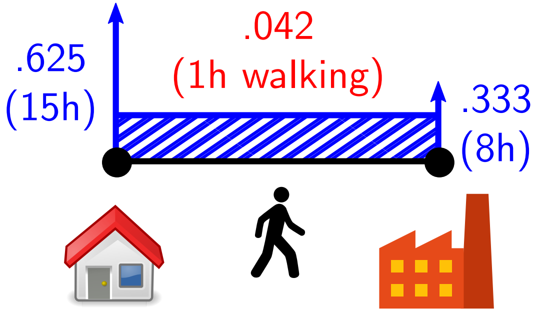
He is in transit 1h/day to commute to work and back.



Back to John's Life

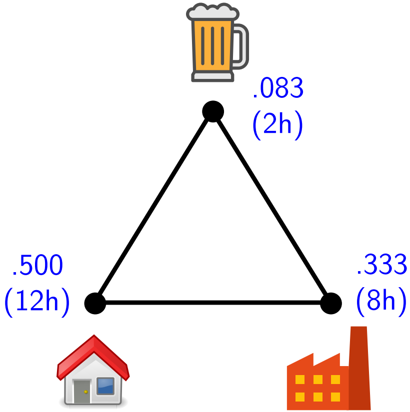
John splits his day as follows: he works 8h/day, and stays home 15h/day.

He is in transit 1h/day to commute to work and back.

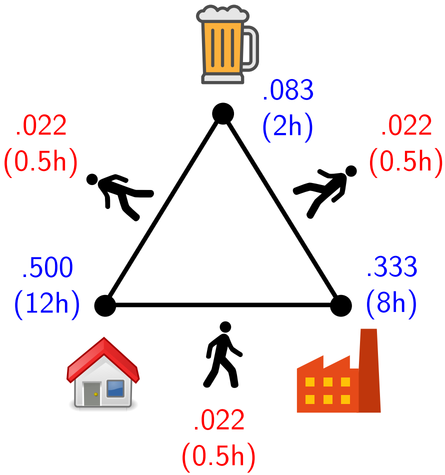


That's a sad life!

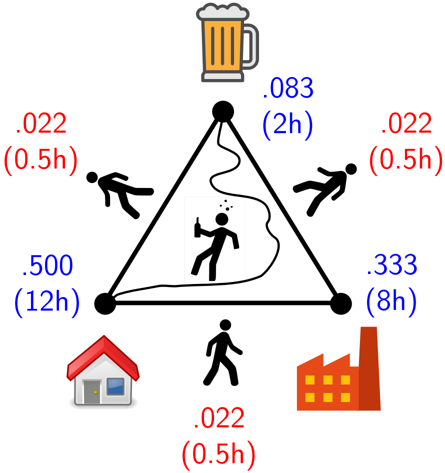
After work, John spends 2h in the pub with friends.



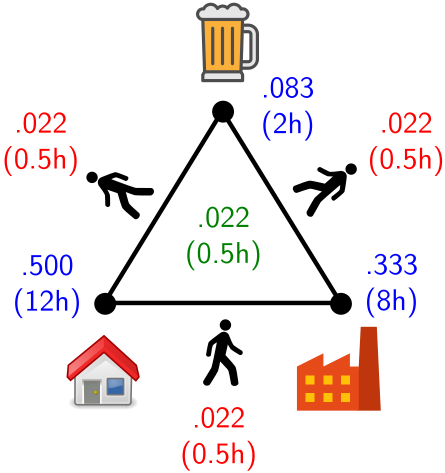
After work, John spends 2h in the pub with friends.



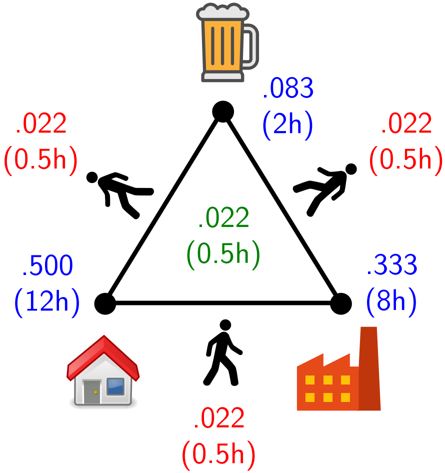
After work, John spends 2h in the pub with friends.



After work, John spends 2h in the pub with friends.



After work, John spends 2h in the pub with friends.



We need a way to represent this probability mass in vertices, edges, face.

Densities over $\triangle_{K-1}$

We denote by $\text{ri}(\triangle_{K-1})$ the relative interior of $\triangle_{K-1}$.

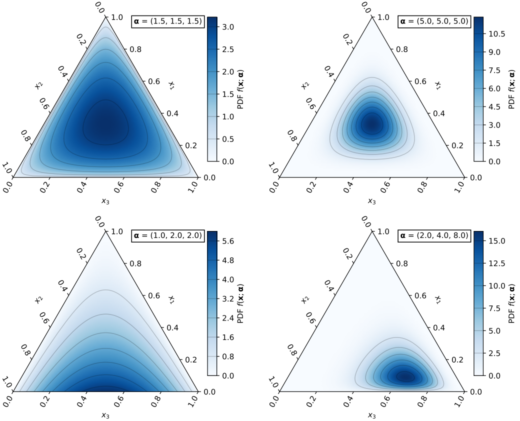
Common densities on the simplex:

- Dirichlet distribution
- Logistic-Normal (a.k.a. Gaussian-Softmax)
- Concrete (a.k.a. Gumbel-Softmax)

None of these place any probability mass on the boundary $\triangle_{K-1} \setminus \text{ri}(\triangle_{K-1})$.

Dirichlet Distribution

$$Y \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}) \quad \Leftrightarrow \quad p_Y(y;\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} y_k^{\alpha_k - 1}, \quad \boldsymbol{\alpha} > 0.$$

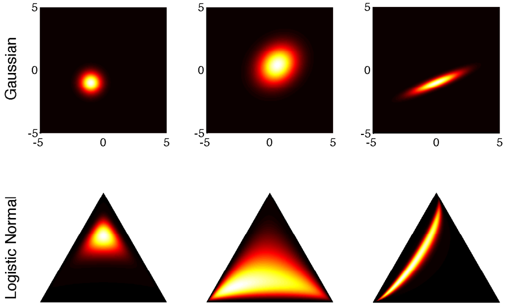


Logistic Normal (a.k.a. Gaussian-Softmax)

(Atchison and Shen, 1980)

Generative story:

$$Y \sim \mathrm{LogisticNormal}(\boldsymbol{z}, \Sigma) \quad \Leftrightarrow \quad \begin{array}{l} N \sim \mathcal{N}(0, \mathrm{I}) \\ Y = \mathrm{softmax}(\boldsymbol{z} + \Sigma^{\frac{1}{2}} N). \end{array}$$



Concrete (a.k.a. Gumbel-Softmax)

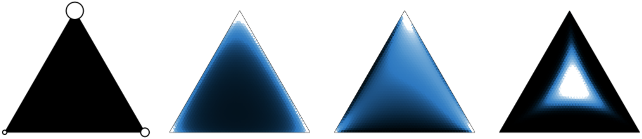
(Maddison et al., 2017; Jang et al., 2017)

Continuous relaxation of a categorical.

Approaches categorical as $\lambda \to 0^+$ (Luce, 1959; Papandreou and Yuille, 2011).

Generative story:

$$Y \sim \text{Concrete}(\boldsymbol{z}, \lambda) \quad \Leftrightarrow \quad \begin{array}{l} G_k \sim \text{Gumbel}(0, 1) \\ Y = \text{softmax}(\lambda^{-1}(\boldsymbol{z} + G)). \end{array}$$



(a) $\lambda = 0$ (b) $\lambda = 1/2$ (c) $\lambda = 1$ (d) $\lambda = 2$

Truncated Densities in the Binary Case ($K = 2$)

When $K = 2$, the simplex is isomorphic to unit interval, $\triangle_1 \simeq [0, 1]$.

A point in $\triangle_1$ can be represented as $\boldsymbol{y} = [y, 1 - y]$.

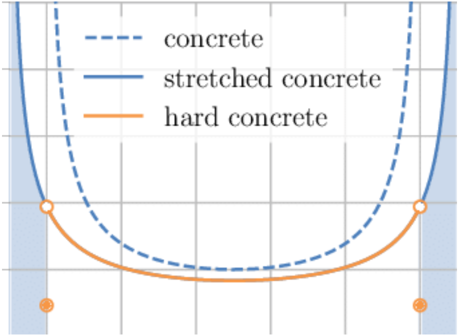
Truncated densities have been proposed for $K = 2$:

- Binary Hard Concrete
- Rectified Gaussian

Binary Hard Concrete

(Louizos et al., 2018)

- Stretches the Concrete and applies a “hard” sigmoid transformation to place point masses at 0 and 1.
- Similar to spike-and-slab (Mitchell and Beauchamp, 1988; Ishwaran et al., 2005).



Rectified Gaussian

(Hinton and Ghahramani, 1997; Palmer et al., 2017)

- Applies a “hard” sigmoid transformation to a univariate Gaussian.

$$p_Y(y) = \mathcal{N}(y; z, \sigma^2) + \frac{1 - \operatorname{erf}\left(\frac{z}{\sqrt{2}\sigma}\right)}{2}\delta_0(y) + \frac{1 + \operatorname{erf}\left(\frac{z-1}{\sqrt{2}\sigma}\right)}{2}\delta_1(y).$$

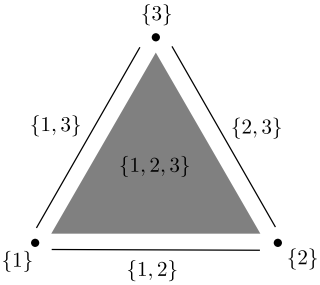
Extending such distributions to the multivariate case ($K > 2$) is non-trivial:

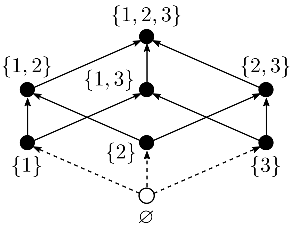
- Combinatorially many multiple order Diracs would be needed
- Dirac deltas have $-\infty$ differential entropy.

Our Approach: Face Stratification

How to extend these “truncated densities” to $K > 2$?

Our solution relies on the face lattice of the simplex:





0-faces are vertices, 1-faces are edges, etc.

There is one $(K-1)$-face: the simplex $\triangle_{K-1}$ itself.

Direct Sum Measure (Farinhas et al., 2022, ICLR)

Let $\mathcal{F}$ denote the set of proper faces of $\triangle_{K-1}$; we have $|\mathcal{F}| = 2^K - 1$.

We define a direct sum measure $\mu^{\oplus}$ on $\triangle_{K-1}$ as a sum of Lebesgue measures on each non-vertex face, and a counting measure on the vertices:

$$\mu^{\oplus}(A) = \sum_{f \in \mathcal{F}} \mu_f(A \cap \mathrm{ri}(f)), \quad A \subseteq \triangle_{K-1}.$$

We define probability densities w.r.t. this base measure.

Mixed Random Variables (Farinhas et al., 2022, ICLR)

Discrete RVs assign probability only to 0-faces (vertices of $\triangle_{K-1}$).

Continuous RVs assign probability only to the maximal face ($\mathrm{ri}(\triangle_{K-1})$).

Mixed RVs generalize both: can assign probability to all faces of $\triangle_{K-1}$.

Mixed Random Variables (Farinhas et al., 2022, ICLR)

Discrete RVs assign probability only to 0-faces (vertices of $\triangle_{K-1}$).

Continuous RVs assign probability only to the maximal face ($\mathrm{ri}(\triangle_{K-1})$).

Mixed RVs generalize both: can assign probability to all faces of $\triangle_{K-1}$.

They can be defined via:

- Their face probability mass function $P_F(f) = \Pr\{\boldsymbol{y} \in \mathrm{ri}(f)\}$, $f \in \mathcal{F}$.
- Their face-conditional densities $p_{Y|F}(\boldsymbol{y} \mid f)$, for $f \in \mathcal{F}$, $\boldsymbol{y} \in \mathrm{ri}(f)$.

The probability of a set $A \subseteq \triangle_{K-1}$ is given by:

$$\Pr\{\boldsymbol{y} \in A\} = \sum_{f \in \mathcal{F}} P_F(f) \int_{A \cap \mathrm{ri}(f)} p_{Y|F}(\boldsymbol{y} \mid f).$$

Extrinsic vs Intrinsic (Farinhas et al., 2022, ICLR)

Two ways of characterizing mixed RVs:

- Extrinsic characterizaton: start with a distribution over $\mathbb{R}^K$ and then apply a deterministic transformation to project it to $\triangle_{K-1}$
- Intrinsic characterizaton: specify a mixture of distributions directly over the faces of $\triangle_{K-1}$, by specifying P_F and $p_{Y|F}$ for each $f \in \mathcal{F}$

K-D Hard Concrete (Farinhas et al., 2022, ICLR)

Uses an extrinsic characterization, via “stretch-and-project.”

Generative story:

$$Y \sim \mathrm{HardConcrete}(\boldsymbol{z}, \lambda, \tau) \quad \Leftrightarrow \quad \begin{array}{l} Y' \sim \mathrm{Concrete}(\boldsymbol{z}, \lambda) \\ Y = \mathrm{sparsemax}(\tau Y'), \quad \text{with } \tau \geq 1. \end{array}$$

- Recovers the binary Hard Concrete for $K = 2$
- The larger τ, the higher the tendency to hit a non-maximal face of the simplex and induce sparsity.

Gaussian-Sparsemax (Farinhas et al., 2022, ICLR)

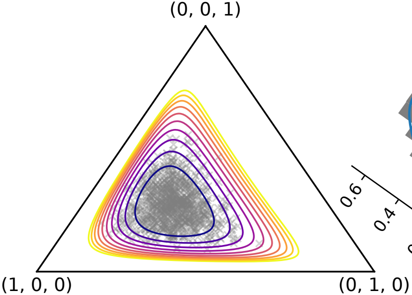
Uses an extrinsic characterization, by sampling from a Gaussian and projecting.

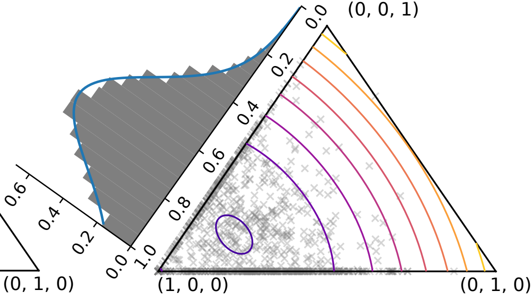
Generative story:

$$Y \sim \mathrm{GaussianSparsemax}(\boldsymbol{z}, \Sigma) \quad \Leftrightarrow \quad \begin{array}{l} N \sim \mathcal{N}(0, \mathrm{I}) \\ Y = \mathrm{sparsemax}(\boldsymbol{z} + \Sigma^{1/2} N). \end{array}$$

- Sparsemax counterpart of the Logistic-Normal.
- Can assign nonzero probability mass to the boundary of the simplex.
- When $K = 2$, we recover the double-sided rectified Gaussian.
- For $K > 2$, an intrinsic representation can be expressed via the orthant probability of multivariate Gaussians.

Logistic-Normal vs Gaussian-Sparsemax (Farinhas et al., 2022, ICLR)





Logistic-Normal (left) assigns zero probability to all faces but $\mathrm{ri}(\triangle_{K-1})$

Gaussian-Sparsemax (right) is a mixed distribution: it assigns probability to the *full* simplex, including its boundary.

Mixed Dirichlet (Farinhas et al., 2022, ICLR)

Uses an intrinsic characterization.

- Uses two parameters: $\boldsymbol{w} \in \mathbb{R}^K$ and $\boldsymbol{\alpha} \in \mathbb{R}^K_{>0}$
- First, sample a face $f \sim P_F(f) \propto \prod_{k \in f} w_k$, where $\boldsymbol{w} \in \mathbb{R}^K$
- Then, sample $Y|F = f \sim \mathrm{Dir}(\boldsymbol{\alpha}|_f)$, where $\boldsymbol{\alpha}|_f$ “masks out” entries of $\boldsymbol{\alpha}$ not supported by f.
- Sampling f can be done in $\mathcal{O}(K)$ with dynamic programming.

Information Theory of Mixed Random Variables

(Farinhas et al., 2022, ICLR)

"Direct sum" entropy using $\mu^{\oplus}$ as the base measure:

$$H^{\oplus}(Y) := H(F) + H(Y \mid F)$$
$$= \underbrace{-\sum_{f \in \mathcal{F}} P_F(f) \log P_F(f)}_{\text{discrete entropy}} + \sum_{f \in \mathcal{F}} P_F(f) \underbrace{\left(-\int_f p_{Y|F}(\boldsymbol{y} \mid f) \log p_{Y|F}(\boldsymbol{y} \mid f) \right)}_{\text{differential entropy}}.$$

- Average length of the optimal code where f must be encoded losslessly and where $\boldsymbol{y}|_f$ has a predefined bit precision N
- Max-ent is written as a generalized Laguerre polynomial (see paper)
  - e.g. $\log_2(2 + 2^N)$ for $K = 2$ (vs. $\log_2(2) = 1$ in the purely discrete case)
- KL divergence and mutual information defined similarly.

Experiment: Emergent Communication

The first agent needs to communicate a code to the second agent that represents a given image.

Given the code, the second agent needs to identify the correct image among 16 possibilities. (Random guess is $1/16 = 6.25\%$.)

Success average and standard error over 10 runs:

Method	Success (%)	Nonzeros ↓
Gumbel-Softmax	78.84 ± 8.07	256
Gumbel-Softmax ST	49.96 ± 9.51	1
K-D Hard Concrete	76.07 ± 7.76	21.43 ± 17.56
Gaussian-Sparsemax	**80.88** ± 0.50	1.57 ± 0.02

(See paper for more experiments with VAEs on FashionMNIST and MNIST.)

Outline

Conclusions

- Transformations from real numbers to distributions are ubiquitous
- We introduced new transformations that handle **sparsity**, **constraints**, and **structure**
- All are differentiable and their gradients are efficient to compute
- Can be used as hidden layers or as output layers (Fenchel-Young losses)
- Mixed distributions are in-between the discrete and continuous worlds
- Examples: Gaussian-Sparsemax, Gumbel-Sparsemax, Mixed Dirichlet
- Sparse communication potentially useful as a path for explainability.

Thank You!

DeepSPIN ("Deep Structured Prediction in NLP")

- ERC starting grant, started in 2018
- Topics: deep learning, structured prediction, NLP
- More details: `https://deep-spin.github.io`

References I

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

Atchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):130–140.

Bastings, J., Aziz, W., and Titov, I. (2019). Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Blondel, M., Martins, A. F. T., and Niculae, V. (2020). Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Correia, G., Niculae, V., and Martins, A. F. T. (2019). Adaptively sparse transformers. In *Proceedings of the Empirical Methods for Natural Language Processing*.

Farinhas, A., Aziz, W., Niculae, V., and Martins, A. F. (2022). Sparse communication via mixed distributions. In *Proc. of ICLR*.

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., and Boyd-Graber, J. (2018). Pathologies of neural models make interpretations difficult. In *Proc. EMNLP*, pages 3719–3728.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

References II

Hinton, G. E. and Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1177–1190.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, 33(2):730–773.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324.

Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. *preprint arXiv:1606.04155*.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2016a). Visualizing and understanding neural models in nlp. In *Proc. NAACL-HLT*, pages 681–691.

Li, J., Monroe, W., and Jurafsky, D. (2016b). Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley, 1959.

References III

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Malaviya, C., Ferreira, P., and Martins, A. F. T. (2018). Sparse and Constrained Attention for Neural Machine Translation. In *Proc. of the Annual Meeting of the Association for Computation Linguistics*.

Martins, A. and Astudillo, R. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA. PMLR.

Martins, A., Farinhas, A., Treviso, M., Niculae, V., Aguiar, P., and Figueiredo, M. (2020a). Sparse and continuous attention mechanisms. *Advances in Neural Information Processing Systems*, 33.

Martins, A. F. T. and Kreutzer, J. (2017). Fully differentiable neural easy-first taggers. In *Proc. of Empirical Methods for Natural Language Processing*.

Martins, P. H., Marinho, Z., and Martins, A. F. (2020b). Sparse text generation. In *Empirical Methods for Natural Language Processing*.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Niculae, V. and Blondel, M. (2017). A regularized framework for sparse and structured neural attention. *arXiv preprint arXiv:1705.07704*.

Niculae, V. and Martins, A. F. (2020). Lp-sparsemap: Differentiable relaxed optimization for sparse structured prediction. In *International Conference on Machine Learning*.

References IV

Niculae, V., Martins, A. F. T., Blondel, M., and Cardie, C. (2018). SparseMAP: Differentiable Sparse Structured Inference. In *Proc. of the International Conference on Machine Learning*.

Palmer, A. W., Hill, A. J., and Scheding, S. J. (2017). Methods for stochastic collection and replenishment (scar) optimisation for persistent autonomy. *Robotics and Autonomous Systems*, 87:51–65.

Papandreou, G. and Yuille, A. L. (2011). Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *2011 International Conference on Computer Vision*, pages 193–200.

Peters, B. and Martins, A. F. (2019). It–ist at the sigmorphon 2019 shared task: Sparse two-headed models for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 50–56.

Peters, B. and Martins, A. F. (2021). Smoothing and shrinking the sparse seq2seq search space. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654.

Peters, B., Niculae, V., and Martins, A. F. T. (2019). Sparse sequence-to-sequence models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proc. ACM SIGKDD*, pages 1135–1144. ACM.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proc. ACL*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

References V

Treviso, M. and Martins, A. F. (2020). The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICLR*, pages 2048–2057.

Zhang, T. (2004). Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *Annals of Statistics*, pages 56–85.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society*, 67(2):301–320.