Sparse Communication via Mixed Distributions

António Farinhas[1] Wilker Aziz[2] Vlad Niculae[3] André F. T. Martins[1,4,5]

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal [2]ILLC, University of Amsterdam, The Netherlands
[3]IvI, University of Amsterdam, The Netherlands [4]Unbabel, Lisbon, Portugal [5]LUMLIS (Lisbon ELLIS Unit), Lisbon, Portugal

Motivation

Commonly we have to opt between *discrete* or *continuous* models

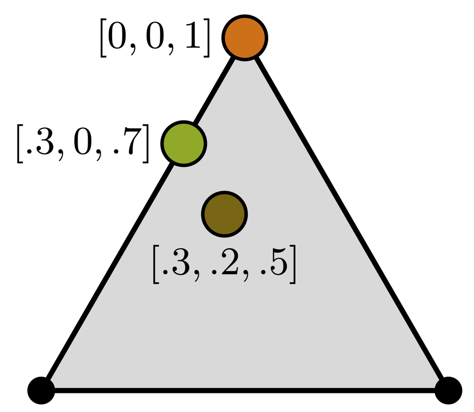
- Language is symbolic and mostly *discrete*
- Neural networks learn and use *continuous* representations

What happens in-between? Can sparsity help?

Transformations from $\mathbb{R}^K$ to $\triangle_{K-1}$

How can we convert a vector of real numbers $z \in \mathbb{R}^K$ (scores for the several symbols, often called *logits*) into a probability vector $y \in \triangle_{K-1}$?

- $y = \mathrm{softmax}(z) \propto \exp(z)$
- $y = \lim_{\tau \to 0^+} \mathrm{softmax}(z/\tau)$
- $y = \mathrm{sparsemax}(z) := \arg\min_{y \in \triangle_{K-1}} \|y - z\|$



Densities over the simplex $\triangle_{K-1}$

There are several works on defining distributions on the simplex $\triangle_{K-1}$

- The Dirichlet, the Concrete (Maddison et al., 2017; Jang et al., 2017), and the Logistic-Normal (Atchison and Shen, 1980) are restricted to $\mathrm{ri}(\triangle_{K-1})$
- The Hard Concrete Louizos et al. (2018) and rectified Gaussians (Palmer et al., 2017) are **mixed** discrete/continuous hybrids limited to $K = 2$

This work:

- Mathematical theory for handling mixed random variables
- Provide extensions to $K > 2$

Extending truncated densities to $K > 2$

We define probability densities with respect to the direct sum measure,

$$\mu^{\oplus}(A) = \sum_{f \in \mathcal{F}} \mu_f(A \cap \mathrm{ri}(f)), \tag{1}$$

where μ_f is the $\dim(f)$-dimensional Lebesgue measure for $\dim(f) > 0$, and the counting measure for $\dim(f) = 0$.

Mixed random variables

How to define probability densities?

1. Define a probability mass function $P_F(f)$ on $\mathcal{F}$
2. For each face $f \in \mathcal{F}$, define a probability density $p_{Y|F}(y \mid f)$ over $\mathrm{ri}(f)$

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

$$\Pr\{y \in A\} = \int_A p_Y^{\oplus}(y) \mathrm{d}\mu^{\oplus} = \sum_{f \in \mathcal{F}} P_F(f) \int_{A \cap \mathrm{ri}(f)} p_{Y|F}(y \mid f) \tag{2}$$

Recovers both discrete and continuous distributions!

Information Theory for mixed random variables

See our paper for generalizations of information theoretic concepts such as entropy, Kullback–Leibler divergence, and mutual information.

The **entropy** of a r.v. X with respect to a measure μ is

$$H^{\mu}(X) = -\int_{\mathcal{X}} p_X(x) \log p_X(x) \mathrm{d}\mu(x), \quad \text{with } \int_{\mathcal{X}} p_X(x) \mathrm{d}\mu(x) = 1 \tag{3}$$

- $\mathcal{X}$ finite, μ counting measure: **Shannon's discrete entropy**
- $\mathcal{X} \subseteq \mathbb{R}^k$ continuous, μ Lebesgue measure: **differential entropy**
- What if μ is the direct sum measure?

$$\begin{aligned} H^{\oplus}(Y) &:= H(F) + H(Y \mid F) \\ &= \underbrace{-\sum_{f \in \mathcal{F}} P_F(f) \log P_F(f)}_{\text{discrete entropy}} + \sum_{f \in \mathcal{F}} P_F(f) \underbrace{\left(-\int_f p_{Y|F}(y \mid f) \log p_{Y|F}(y \mid f) \right)}_{\text{differential entropy}} \end{aligned} \tag{4}$$

Average length of the optimal code where the sparsity pattern of $y \in \triangle_{K-1}$ must be encoded losslessly and where there is a predefined bit precision for the fractional entries of y.

The maximum entropy mixed distribution is written as a generalized Laguerre polynomial

- $\log_2(2 + 2^N)$ for $K = 2$, instead of $\log_2(2) = 1$ in the purely discrete case

Intrinsic characterization

Specify a mixture of distributions directly over the faces of $\triangle_{K-1}$: P_F and $p_{Y|F}$ for each $f \in \mathcal{F}$.

Mixed Dirichlet (two parameters: $w \in \mathbb{R}^K$ and $\alpha \in \mathbb{R}^K_{>0}$)

- Sample a face $f \sim P_F(f) \propto \prod_{k \in f} \exp(w_k)$
- Sample $Y|F = f \sim \mathrm{Dir}(\alpha|_f)$ where $\alpha|_f$ masks out entries of α not supported by f

Extrinsinc characterization

Start with a distribution over the ambient space and project it to the simplex using sparsemax.

K-D Hard Concrete (generalization of the binary Hard Concrete for $K > 2$)

$$Y' \sim \mathrm{Concrete}(z, \beta), \quad Y = \mathrm{sparsemax}(\lambda Y'), \quad \text{with } \lambda \geq 1. \tag{5}$$

Gaussian-Sparsemax (generalization of a double-sided rectified Gaussian for $K > 2$)

$$N \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad Y = \mathrm{sparsemax}(z + \Sigma^{1/2} N) \tag{6}$$

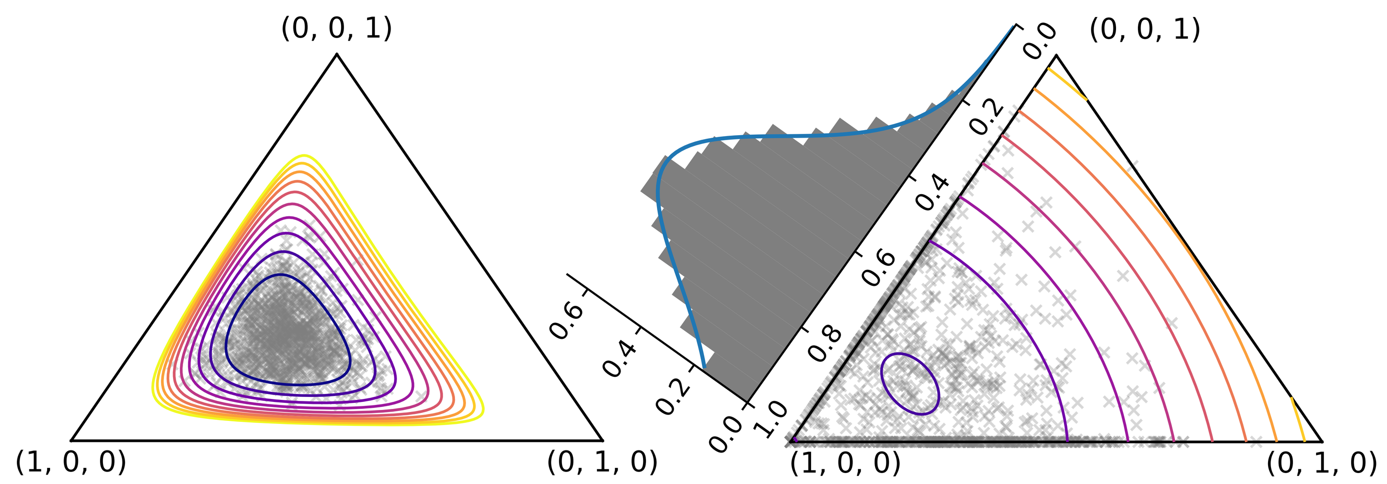


Figure: The Logistic-Normal (left), assigns zero probability to all faces but $\mathrm{ri}(\triangle_{K-1})$. The Gaussian-Sparsemax (right) induces a distribution over the 1-dimensional edges (shown as a histogram), and assigns $\Pr\{(1, 0, 0)\} = .022$.

Can also be defined intrinsically

- Expressed via the orthant probability of multivariate Gaussians for $K > 2$
- Simple expression for $K = 2$; entropy and KL divergence in closed form

Experiments

Emergent communication game (inducing sparse communication between two agents)

- A *sender* sees an image and emits a single-symbol message from a fixed vocabulary
- A *receiver* reads the symbol and tries to identify the correct image out of a set of 16

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

Bit-Vector VAE on Fashion-MNIST (studying the impact of the direct sum entropy)

- We consider 128 binary latent bits and maximize the ELBO

Method	Entropy	NLL	Sparsity (%) ↑	
Binary Concrete	C ≈	3.60	0	
Gumbel-Softmax	D =	**3.49**	0	**Incoherent (discrete) entropy computation**
Gumbel-Softmax ST	D =	3.57	100	
Hard Concrete	X ≈	3.57	45.64	
Gaussian-Sparsemax	X ≈	3.53	82.82	**Coherent objective and unbiased gradients!**
Gaussian-Sparsemax	X =	**3.49**	73.83	

Regression towards voting proportions (using the Mixed Dirichlet as a likelihood function)

- We use the UK election data; the observations are vectors of proportions over 5 parties
- Modeling simplex-valued data with the Dirichlet is tricky

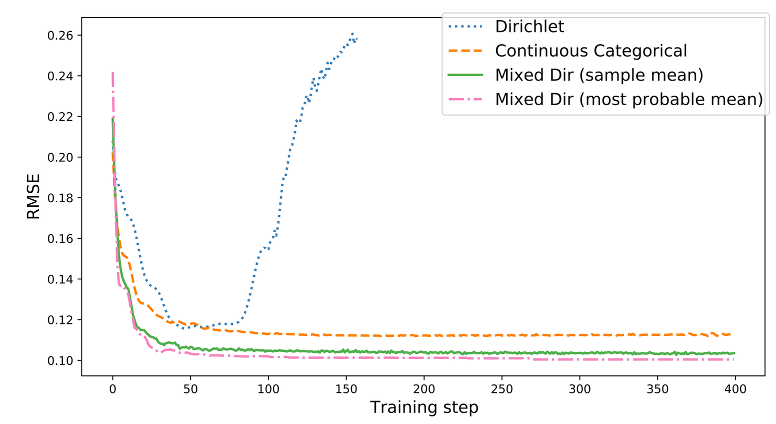


Figure: The Mixed Dirichlet addresses the pathologies of the Dirichlet in this setting, showing a slight advantage over the continuous categorical (Gordon-Rodriguez et al., 2020), likely due to the fact that Mixed Dirichlet samples are often sparse at test time.

Conclusions

- Mathematical framework for handling mixed random variables
- Direct sum measure as an alternative to the Lebesgue-Borel and the counting measures
- Generalizations of information theoretic concepts for mixed symbols
- Experiments on learning sparse representations and avoiding ill-defined log-likelihoods
- **Future work:** More effective intrinsic parametrizations; mixed *structured* variables

Open-source code: https://github.com/deep-spin/sparse-communication

References

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

Gordon-Rodriguez, E., Loaiza-Ganem, G., and Cunningham, J. (2020). The continuous categorical: a novel simplex-valued exponential family. In *Proceedings of ICML*.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In *Proceedings of ICLR*.

Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l_0 regularization. In *Proceedings of ICLR*.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of ICLR*.

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*.

Acknowledgments: P2020 program MAIA, ERC StG DeepSPIN, FCT Project PRELUNA, European Union's Horizon 2020 research and innovation programme GoURMET, Hybrid Intelligence Centre.