Sparse and Structured Hopfield Networks

Saul Santos[1] Vlad Niculae[2] Daniel McNamee[3] André F. T. Martins[1,4]

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon ELLIS Unit, Portugal
[2]Informatics Institute, University of Amsterdam, The Netherlands
[3]Neuroscience Programme, Champalimaud Research, Lisbon, Portugal [4]Unbabel, Lisbon, Portugal

Outline

We extend **modern Hopfield networks** (MHNs) [1] and their sparse variants [2, 3] to a broader family of energy functions, via **Fenchel-Young losses** [4].

- Still end-to-end differentiable, but allow for **exact convergence** to single memory patterns and **exponential storage capacity**.
- Extension to structures, allowing retrieval of pattern associations, via **SparseMAP** [5].
- Experiments on synthetic and real word data (multiple instance learning and text rationalization).

Fenchel-Young Losses [4]

Let $\Omega : \triangle \to \mathbb{R}$ be a **convex regularizer** ($\triangle \equiv$ simplex).

- **Ω-regularized prediction map:**

$$\hat{y}_\Omega(\theta) = \arg\max_{y \in \triangle} \theta^\top y - \Omega(y).$$

- **Fenchel-Young loss** induced by Ω:

$$L_\Omega(\theta, y) = \Omega(y) + \Omega^*(\theta) - \theta^\top y.$$

Examples:

- **Shannon negentropy:** $\Omega(y) = \sum_i p_i \log p_i$
  $\Rightarrow$ softmax & cross-entropy loss
- **Tsallis α-negentropies** [6] with $\alpha \geq 1$: $\Omega^T_\alpha(y) = \frac{-1+\|y\|^\alpha_\alpha}{\alpha(\alpha-1)}$
  $\Rightarrow$ **α-entmax transformations & losses** [7]
- **Norm α-negentropies:** $\Omega^N_\alpha(y) = -1 + \|y\|_\alpha$
  $\Rightarrow$ **α-normmax transformations & losses** [4].

Properties:

- $L_\Omega(\theta, y) \geq 0$, with equality iff $y = \hat{y}_\Omega(\theta)$.
- $L_\Omega(\theta, y)$ is convex on θ and $\nabla_\theta L_\Omega(\theta, y) = -y + \hat{y}_\Omega(\theta)$.

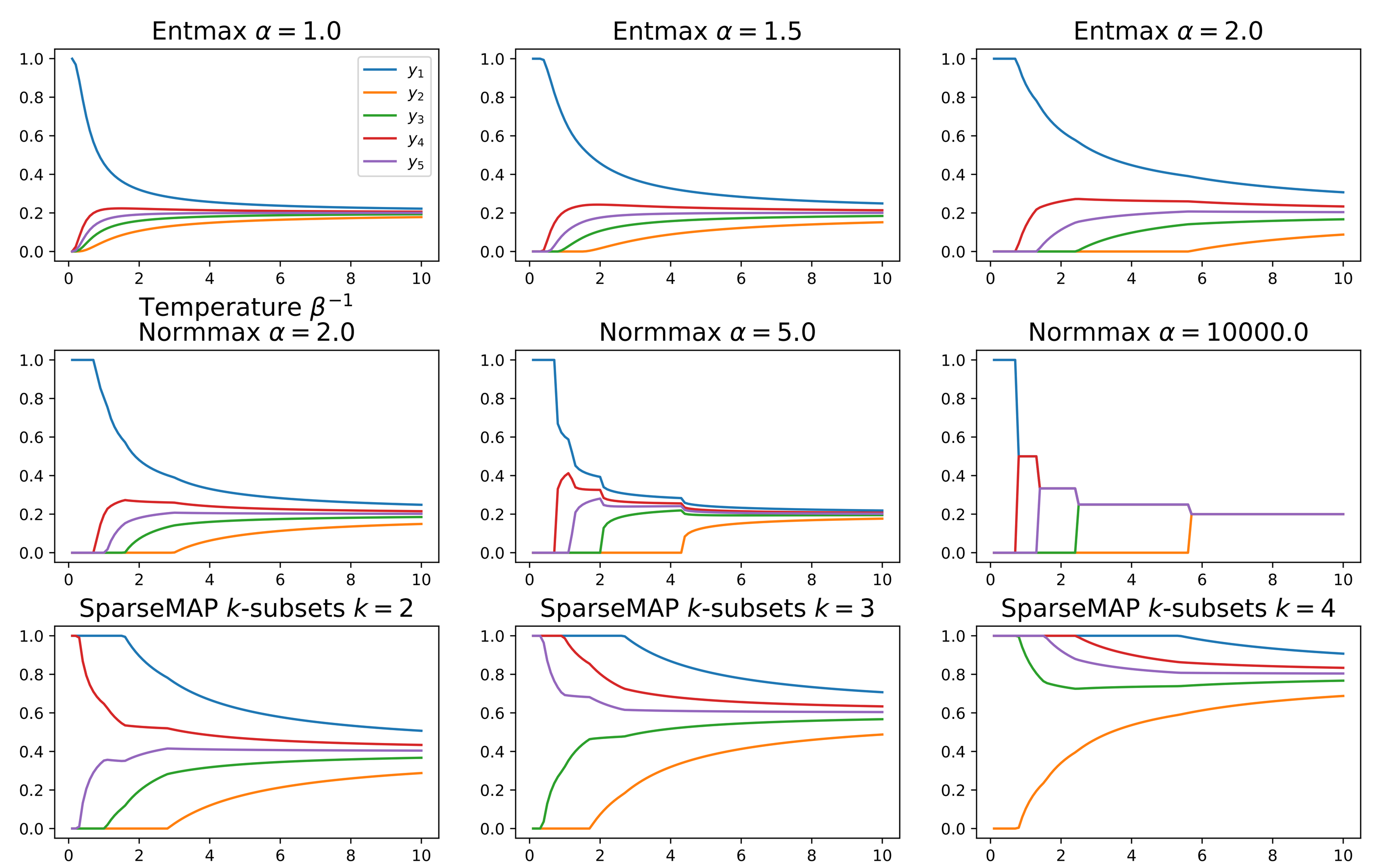
Margin Property

L_Ω has the **margin property** with margin $m > 0$ if:

$$L_\Omega(\theta, e_i) = 0 \;\Leftrightarrow\; \hat{y}_\Omega(\theta) = e_i \;\Leftrightarrow\; \theta_i - \max_{j \neq i} \theta_j \geq m.$$

For α-entmax, $m = 1/(\alpha - 1)$; for α-normmax, $m = 1$.

Sparse and Structured Transformations



This Paper: Sparse Hopfield Networks

Set of N memory patterns $X \in \mathbb{R}^{N \times D}$, query $q \in \mathbb{R}^D$

- **Hopfield-Fenchel-Young energy:**

$$E(q) = \underbrace{-\beta^{-1} L_\Omega(\beta X q; \mathbf{1}/N)}_{E_{\text{concave}}(q)} + \underbrace{\frac{1}{2}\|q - X^\top \mathbf{1}/N\|^2 + \text{const.}}_{E_{\text{convex}}(q)}.$$

- **Update rule** (via CCCP):

$$q_{t+1} = X^\top \hat{y}_\Omega(\beta X q_t).$$

Subsumes MHNs [1] and sparse variants [2, 3].

Exact Convergence and Exponential Memory Capacity

Separation of pattern x_i from data: $\Delta_i = \min_{j \neq i} x_i^\top (x_i - x_j)$

Proposition: Assume L_Ω has margin m. Then:

- x_i is a stationary point of the HFY energy iff $\Delta_i \geq m\beta^{-1}$
- If the patterns are normalized (radius M) and $\Delta_i \geq m\beta^{-1} + 2M\epsilon$, then any q_0 ϵ-close to x_i ($\|q_0 - x_i\| \leq \epsilon$) will converge to x_i in 1 iteration.
- With probability $1 - p$, the HFY network can store and exactly retrieve $N = \mathcal{O}(\sqrt{p}\zeta^{\frac{D-1}{2}})$ patterns in 1 iteration under a ϵ-perturbation if $\epsilon \leq \frac{M}{2}\left(1 - \cos\frac{1}{\zeta}\right) - \frac{m}{2\beta M}$.

This paper: Structured Hopfield Networks

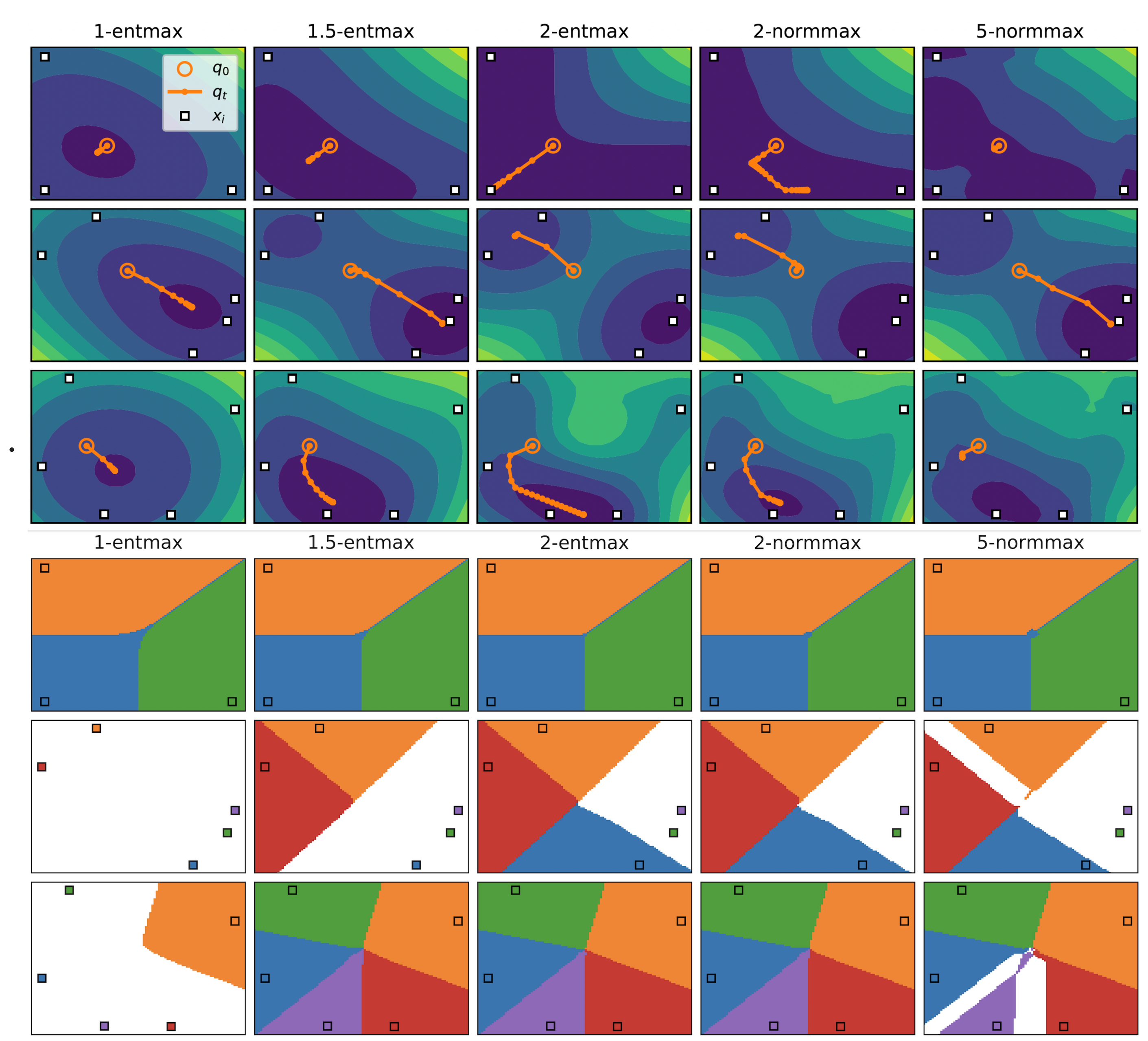
FY structured losses replace $\triangle$ by marginal polytope representing a structured space.

- **k-subsets**: Promotes top-k retrieval.
- **sequential k-subsets**: Promotes consecutive memory items to be both retrieved or neither retrieved.

This can be accomplished with **SparseMAP** with **exact structured retrieval** (see paper):

- Hopfield dynamics $q_{t+1} = X^\top \text{SparseMAP}(\beta X q)$

Hopfield Dynamics and Basins of Attraction



K-MIL on MNIST

Methods	K=2	K=3	K=5
SparseMAP, $k = 2$	**97.7 ± 0.3**	95.1 ± 0.5	92.6 ± 1.1
SparseMAP, $k = 3$	96.1 ± 1.0	**96.5 ± 0.5**	92.2 ± 1.2
SparseMAP, $k = 5$	96.2 ± 1.4	95.1 ± 1.1	**95.1 ± 1.5**

Structured Rationalizers

Sequential k-subsets

a darkish golden pour from tap with a small white lacing around glass . you can't miss the sweet smell . the word snappy fits this beer well . it is a winter warmer but not from the usual alcohol burn . the alcohol is almost completely hidden . the warm comes from the mix of cinnamon , hops , and most of all spiciness . the alcohol must be there because i sure did feel it after finishing the glass .

k-subsets

a darkish golden pour from tap with a small white lacing around glass . you can't miss the sweet smell . the word snappy fits this beer well . it is a **winter warmer** but not from the usual alcohol burn . the alcohol is almost completely hidden . the warm comes from the mix of **cinnamon** , hops , and most of all **spiciness** . the alcohol must be there because i sure did feel it after finishing the **glass** .

Rationales from our Hopfield pooling layer: sparseMAP generator with k-subsets and sequential k-subsets.

	AgNews↑	Beer (MSE) ↓	Beer (HRO) ↑
HardKuma [8]	.90 (.87/.88)	.019 (.016/.020)	.37 (.00/.90)
SPECTRA [9]	.92 (.92/.93)	**.017** (.016/.019)	.61 (.56/.68)
SparseMAP k-subsets (ours)	**.93** (.92/.93)	**.017** (.017/.018)	.42 (.29/.62)
SparseMAP seq. k-subsets (ours)	**.93** (.93/.93)	.020 (.018/.021)	**.63** (.49/.70)

Text rationalization results. We report mean and min/max MSE for beer and F_1 scores for AgNews across five random seeds.

References

[1] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *Proceedings of ICLR*, 2021.

[2] Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. On sparse modern hopfield model. In *NeurIPS*, 2023.

[3] Dennis Wu, Jerry Yao-Chieh Hu, Weijian Li, Bo-Yu Chen, and Han Liu. STanhop: Sparse tandem hopfield model for memory-enhanced time series prediction. In *Proceedings of ICLR*, 2024.

[4] Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(1):1314–1382, 2020.

[5] Vlad Niculae, Andre Martins, Mathieu Blondel, and Claire Cardie. Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning*, pages 3799–3808. PMLR, 2018.

[6] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

[7] Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Proceedings of ACL*, 2019.

[8] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of ACL*, pages 2963–2977, 2019.

[9] Nuno M. Guerreiro and André F. T. Martins. Spectra: Sparse structured text rationalization. In *Proceedings of EMNLP*, pages 6534–6550, 2021.

Acknowledgments: ERC (DECOLLAGE, ERC-2022-CoG 101088763), Portuguese PRR (Center for Responsible AI, C645008882- 00000055), FCT (contract UIDB/50008/2020), Dutch Research Council (NWO) via VI.Veni.212.228.