Sparse Modern Hopfield Networks

André F. T. Martins
Instituto de Telecomunicações
Instituto Superior Técnico (Lisbon ELLIS Unit) & Unbabel
Lisbon, Portugal
andre.t.martins@tecnico.ulisboa.pt

Vlad Niculae
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
v.niculae@uva.nl

Daniel McNamee
Neuroscience Programme
Champalimaud Research
Lisbon, Portugal
daniel.mcnamee@research.fchampalimaud.org

Abstract

Ramsauer et al. (2021) recently pointed out a connection between modern Hopfield networks and attention heads in transformers. In this paper, we extend their framework to a broader family of energy functions which can be written as a difference of a quadratic regularizer and a Fenchel-Young loss (Blondel et al., 2020), parametrized by a generalized negentropy function Ω. By working with Tsallis negentropies, the resulting update rules become end-to-end differentiable *sparse* transformations, establishing a new link to adaptively sparse transformers (Correia et al., 2019) and allowing for exact convergence to single memory patterns. Experiments on simulated data show a higher tendency to avoid metastable states.

1 Introduction

Hopfield networks are a kind of biologically-plausible neural network exhibiting associative memory capabilities (Hopfield, 1982). Their attractor dynamics makes them suitable for modeling the retrieval of episodic memories in humans and animals (Tyulmankov et al., 2021; Whittington et al., 2021). The limited storage capabilities of classical (quadratic energy) Hopfield networks were recently overcome through new energy functions and continuous state patterns (Krotov and Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2021), leading to exponential storage capacities, and sparking renewed interest in modern Hopfield networks. In particular, Ramsauer et al. (2021) revealed striking connections to transformer attention, via an update rule linked to the convex-concave procedure (CCCP; Yuille and Rangarajan 2003). However, this model has the often-undesirable tendency to converge to large metastable states (mixing many input patterns) instead of retrieving a single pattern.

There is a strong neurobiological motivation to seek new Hopfield energies capable of sparse selection of patterns. Sparse neural activity patterns are observed in electrophysiological recordings from many brain areas across a variety of animal species and forms a core principle of cortical computation due to their efficient coding properties (Simoncelli and Olshausen, 2001; Palm, 2013). With respect to memory formation circuits, the sparse firing of neurons in the dentate gyrus (DG), a major input pathway to the CA3 and CA1 subregions in the hippocampus, underpins its theorized role in pattern separation during memory storage (Yassa and Stark, 2011; Severa et al., 2017). Indeed, evidence suggests that the sparsified activity profiles of DG neurons aids in minimizing interference between competing memory patterns just prior to pattern completion via autoassociative dynamics downstream (Leutgeb et al., 2007; Neunuebel and Knierim, 2014).

Associative Memory & Hopfield Networks in 2023. NeurIPS 2023 workshop.

In this paper, inspired by the framework of regularized prediction maps and Fenchel-Young losses (Blondel et al., 2020), we extend Ramsauer et al.'s (2021) energy function to a wider family induced by generalized entropies. The minimization of these energy functions leads to update rules which include as particular cases **sparsemax** (Martins and Astudillo, 2016) and the α**-entmax transformations** used by adaptively sparse transformers (Peters et al., 2019; Correia et al., 2019). Unlike Ramsauer et al.'s (2021) Hopfield layers, our proposed update rules can lead to **sparse** convex combinations of input patterns and can have *exact* convergence to a fixed point in a small number of steps, maintaining end-to-end differentiability. Experiments on simulated data show frequent convergence to a *single* input pattern or small metastable states.

2 Modern Hopfield Networks

Let $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ be a matrix whose rows hold a set of examples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ ("memory patterns"), where each $\boldsymbol{x}_i \in \mathbb{R}^D$, and let $\boldsymbol{q}_0 \in \mathbb{R}^D$ be a query vector (or "state pattern"). The Hopfield network iteratively updates $\boldsymbol{q}_t \mapsto \boldsymbol{q}_{t+1}$ for $t \in \{0, 1, \ldots\}$ according to a certain rule and, under certain conditions, these dynamical trajectories converge to a fixed point attractor state $\boldsymbol{q}^*$ which either corresponds to one of the memorized examples, or to a mixture thereof. This update rule correspond to the minimization of an energy function, which for classic Hopfield networks (Hopfield, 1982) takes the form $E(\boldsymbol{q}) = -\frac{1}{2}\boldsymbol{q}^\top \boldsymbol{W}\boldsymbol{q}$, where $\boldsymbol{W} = \boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{D \times D}$ are the Hopfield network parameters, which when $D \ll N$ can be seen as a "compressed memory". In the classic Hopfield network, the state vector $\boldsymbol{q}$ is further constrained to be $\{\pm 1\}$-valued, and the update rule is $\boldsymbol{q}_{t+1} = \mathrm{sign}(\boldsymbol{W}\boldsymbol{q}_t)$. A limitation of this classical network is that it has only $O(D)$ memory storage capacity.

Recent work sidestepped this limitation through alternative energy functions (Krotov and Hopfield, 2016; Demircigil et al., 2017), leading to the development of a class of models known as "modern Hopfield networks" with superlinear (often exponential) memory capacity. In Ramsauer et al. (2021), the state vector $\boldsymbol{q} \in \mathbb{R}^D$ is continuous and unconstrained and the following energy is used:

$$E(\boldsymbol{q}) = -\mathrm{lse}(\beta, \boldsymbol{X}\boldsymbol{q}) + \frac{1}{2}\boldsymbol{q}^\top \boldsymbol{q} + \beta^{-1}\log N + \frac{1}{2}M^2, \tag{1}$$

where $M = \max_i \|\boldsymbol{x}_i\|$ and $\mathrm{lse}(\beta, \boldsymbol{\theta}) = \beta^{-1}\log\sum_{i=1}^N \exp(\beta\theta_i)$.

Ramsauer et al. (2021) have revealed an interesting relation between the updates in this modern Hopfield network and the attention layers in transformers. Namely, the minimization of the energy (1) using the concave-convex procedure (CCCP; Yuille and Rangarajan 2003) leads to the update rule

$$\boldsymbol{q}_{t+1} = \boldsymbol{X}^\top \mathrm{softmax}(\beta \boldsymbol{X}\boldsymbol{q}_t). \tag{2}$$

When $\beta = \frac{1}{\sqrt{D}}$, each update matches the computation performed in the attention layer of a transformer with a single attention head and with identity projection matrices. This triggered interest in developing variants of Hopfield layers which can be used as drop-in replacements for multi-head attention layers.

While Ramsauer et al. (2021) have derived useful theoretical properties of these networks (including their exponential storage capacity under some assumptions), the use of softmax in the update rule (2) prevents exact convergence and may lead to undesirable metastable states in some situations, as illustrated in toy experiments in §5. We explore in this paper a more general energy function which uses sparse transformations as an attempt to overcome these drawbacks.

3 Ω-Regularized Prediction Maps and Fenchel-Young Losses

Our contribution is rooted in the concept of Fenchel-Young losses, which we next review. Let $\triangle_{N-1} := \{\boldsymbol{p} \in \mathbb{R}^N : \boldsymbol{p} \geq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{p} = 1\}$ denote the probability simplex, whose elements are probability vectors of length N. Given a convex function $\Omega : \triangle_{N-1} \to \mathbb{R}$, the Ω**-regularized prediction map** (Ω-RPM; Blondel et al. 2020), $\hat{\boldsymbol{p}}_\Omega : \mathbb{R}^N \to \triangle_{N-1}$, is:

$$\hat{\boldsymbol{p}}_\Omega(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{p} \in \triangle_{N-1}} \boldsymbol{\theta}^\top \boldsymbol{p} - \Omega(\boldsymbol{p}). \tag{3}$$

One example of an Ω-RPM is the **softmax** transformation, obtained when $\Omega(\boldsymbol{p}) = \sum_{i=1}^N p_i \log p_i$ is the Shannon negentropy. Another example is the **sparsemax** transformation, obtained when $\Omega(\boldsymbol{p}) =$

$\frac{1}{2}\|\boldsymbol{p}\|^2$ (Martins and Astudillo, 2016). The sparsemax corresponds to the Euclidean projection onto the probability simplex. Softmax and sparsemax are both particular cases of α**-entmax transformations** (Peters et al., 2019), parametrized by a scalar $\alpha \geq 0$ (called the entropic index), which correspond to the following choice of regularizer, called the **Tsallis** α**-negentropy** (Tsallis, 1988):

$$\Omega_\alpha(\boldsymbol{p}) = \left(-1 + \|\boldsymbol{p}\|_\alpha^\alpha\right)/\alpha(\alpha - 1). \tag{4}$$

Note that, when $\alpha \to 1$, Ω_α becomes Shannon's negentropy and the corresponding Ω-RPM is the softmax, and when $\alpha = 2$, it becomes the ℓ_2-norm (up to a constant) and we recover the sparsemax.

Let Ω^* be the convex conjugate of Ω, $\Omega^*(\boldsymbol{\theta}) = \max_{\boldsymbol{p}\in\triangle_{N-1}} \boldsymbol{\theta}^\top \boldsymbol{p} - \Omega(\boldsymbol{p})$. The Ω-RPM in (3) equals the gradient map of Ω^*, $\hat{p}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta})$. Note also that we have $\Omega^*(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \hat{\boldsymbol{p}}_\Omega(\boldsymbol{\theta}) - \Omega(\hat{\boldsymbol{p}}_\Omega(\boldsymbol{\theta}))$. The **Fenchel-Young loss** induced by Ω (Blondel et al., 2020) is the function defined as

$$L_\Omega(\boldsymbol{\theta}, \boldsymbol{p}) = \Omega(\boldsymbol{p}) + \Omega^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{p}. \tag{5}$$

When Ω is the Shannon negentropy, $\Omega^*(\boldsymbol{\theta}) = \mathrm{lse}(1, \boldsymbol{\theta})$, and L_Ω is the **cross-entropy loss**, up to a constant (Blondel et al., 2020, §3.2). Intuitively, Fenchel-Young losses quantify how "compatible" a score vector $\boldsymbol{\theta} \in \mathbb{R}^N$ (e.g., logits) is to a desired probability vector $\boldsymbol{p} \in \triangle_{N-1}$. Additionally, (Blondel et al., 2020, Prop. 2):

1. Fenchel-Young losses are non-negative, $L_\Omega(\boldsymbol{\theta}, \boldsymbol{p}) \geq 0$, with equality iff $\boldsymbol{p} = \hat{\boldsymbol{p}}_\Omega(\boldsymbol{\theta})$.
2. Fenchel-Young losses are convex on $\boldsymbol{\theta}$ and their gradient is $\nabla_{\boldsymbol{\theta}} L_\Omega(\boldsymbol{\theta}, \boldsymbol{p}) = -\boldsymbol{p} + \hat{\boldsymbol{p}}_\Omega(\boldsymbol{\theta})$.

For Tsallis negentropies Ω_α with $\alpha > 1$, a margin property holds (Blondel et al., 2020, Prop. 7):

$$\forall i \in [N], \quad L_{\Omega_\alpha}(\boldsymbol{\theta}, \boldsymbol{e}_i) = 0 \iff \hat{p}_{\Omega_\alpha}(\boldsymbol{\theta}) = \boldsymbol{e}_i \iff \theta_i - \max_{j\neq i} \theta_j \geq (\alpha - 1)^{-1}. \tag{6}$$

We will use these properties in the next section to define and analyze sparse Hopfield networks.

4 Sparse Hopfield Networks

We now use Ω-RPMs and Fenchel-Young losses to define a new class of energy functions associated to modern Hopfield networks. We assume that the regularizer Ω is a **generalized negentropy**, i.e., null when $\boldsymbol{p}$ is peaked, strictly convex, and permutation-invariant (see Appendix A). These conditions imply that $\Omega \leq 0$ and that Ω is minimized when $\boldsymbol{p} = \mathbf{1}/N$ is the uniform distribution (Blondel et al., 2020, Prop. 4). The Tsallis negentropy (4) satisfies these properties for $\alpha \geq 1$.

We define the **Hopfield-Fenchel-Young energy** as

$$E(\boldsymbol{q}) = \underbrace{-\beta^{-1} L_\Omega(\beta \boldsymbol{X}\boldsymbol{q}; \mathbf{1}/N)}_{E_{\mathrm{concave}}(\boldsymbol{q})} + \underbrace{\frac{1}{2}\|\boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}}\|^2 + \frac{1}{2}(M^2 - \|\boldsymbol{\mu}_{\boldsymbol{X}}\|^2)}_{E_{\mathrm{convex}}(\boldsymbol{q})}, \tag{7}$$

where $\boldsymbol{\mu}_{\boldsymbol{X}} := \boldsymbol{X}^\top \mathbf{1}/N \in \mathbb{R}^D$ is the empirical mean of the patterns. This energy extends that of (1), which is recovered when Ω is Shannon's negentropy, in which case $E_{\mathrm{concave}}(\boldsymbol{q}) = -\mathrm{lse}(\beta, \boldsymbol{X}\boldsymbol{q}) + \beta^{-1}\log N + \boldsymbol{q}^\top \boldsymbol{\mu}_{\boldsymbol{X}}$. The concavity of E_{concave} holds from the convexity of Fenchel-Young losses on its first argument and from the fact that composition of a convex function with an affine map is convex. The convexity of E_{convex} comes from the fact that it is a quadratic function.[1]

There are two terms competing when minimizing the energy function (7) with respect to $\boldsymbol{q}$:

- Minimizing E_{concave} is equivalent to *maximizing* $L_\Omega(\beta \boldsymbol{X}\boldsymbol{q}; \mathbf{1}/N)$, which pushes for state patterns $\boldsymbol{q}$ as far from possible from a uniform average and close to a single memory pattern.
- Minimizing E_{convex} serves as a regularization, encouraging the state pattern to stay close to $\boldsymbol{\mu}_{\boldsymbol{X}}$.

The next result, proved in Appendix B, establishes bounds and derives the Hopfield update rule for our energy function, generalizing Ramsauer et al. (2021, Lemma A.1 and Theorem A.1).

[1] Up to constants, for Ω_1 this is the same convex-concave decomposition of Ramsauer et al. (2021)

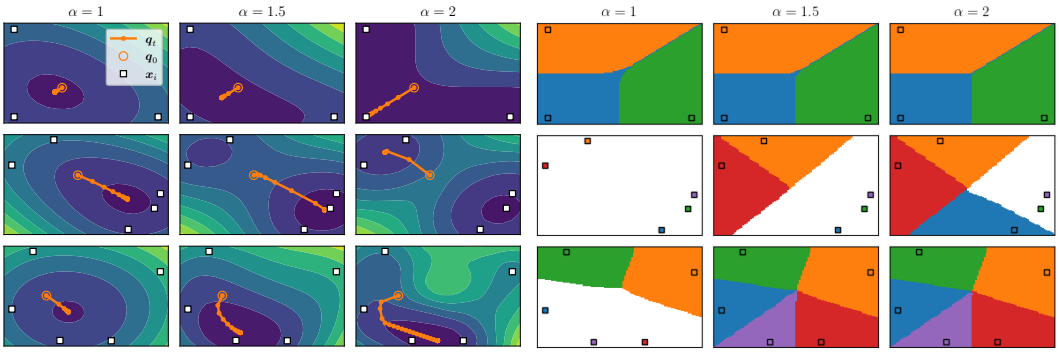


Figure 1: Left: contours of the energy function and optimization trajectory of the CCCP iteration ($\beta = 1$). Right: attraction basins associated with each pattern. (White sections do not converge to a single pattern but to a metastable state; $\beta = 10$; for $\alpha = 1$ we allow a tolerance of $\epsilon = .01$) .

Proposition 1. *Let the query $\boldsymbol{q}$ be in the convex hull of the rows of $\boldsymbol{X}$, i.e., $\boldsymbol{q} = \boldsymbol{X}^\top \boldsymbol{p}$ for some $\boldsymbol{p} \in \triangle_{N-1}$. Then, the energy* (7) *satisfies $0 \leq E(\boldsymbol{q}) \leq \min\left\{2M^2,\ -\beta^{-1}\Omega(\mathbf{1}/N) + \frac{1}{2}M^2\right\}$. Furthermore, minimizing* (7) *by CCCP (Yuille and Rangarajan, 2003) leads to the updates:*

$$\boldsymbol{q}_{t+1} = \boldsymbol{X}^\top \hat{\boldsymbol{p}}_\Omega(\beta \boldsymbol{X} \boldsymbol{q}_t). \tag{8}$$

In particular, when $\Omega = \Omega_\alpha$ (the Tsallis α-entropy (4)*), the Ω-RPM is the α-entmax transformation, corresponding to the adaptively sparse transformer of Correia et al. (2019).*

We now show that, with $\Omega = \Omega_\alpha$ and for $\alpha > 1$ (the sparse case), the memory patterns can be stationary points of the energy (7). This result is stronger than that of Ramsauer et al. (2021) for their energy (which is ours for $\alpha = 1$), according to which memory patterns are only ϵ-close to stationary points, where a small $\epsilon = O(\exp(-\beta))$ requires a low temperature (large β). The proof (Appendix C) relies on the margin property stated in (6). Following Ramsauer et al. (2021, Def. 2), we define the separation of pattern $\boldsymbol{x}_i$ from data as $\Delta_i = \boldsymbol{x}_i^\top \boldsymbol{x}_i - \max_{j \neq i} \boldsymbol{x}_i^\top \boldsymbol{x}_j$.

Proposition 2. *Assume $\Omega = \Omega_\alpha$ with $\alpha > 1$, and let $\boldsymbol{x}_i$ be a memory pattern outside the convex hull of the other memory patterns. Then, $\boldsymbol{x}_i$ is a stationary point of the energy* (7) *iff $\Delta_i \geq \frac{1}{(\alpha-1)\beta}$. In addition, if the initial query $\boldsymbol{q}_0$ satisfies $\boldsymbol{q}_0^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq \frac{1}{(\alpha-1)\beta}$ for all $j \neq i$, then the update rule* (8) *converges to $\boldsymbol{x}_i$ exactly in one iteration. Moreover, if the patterns are normalized and $\Delta_i \geq \frac{1}{(\alpha-1)\beta} + 2\epsilon$, then any $\boldsymbol{q}_0$ ϵ-close to $\boldsymbol{x}_i$ ($\|\boldsymbol{q}_0 - \boldsymbol{x}_i\| \leq \epsilon$) will converge to $\boldsymbol{x}_i$ in one iteration.*

We next validate these results on simulated data.

5 Experiments

We repeatedly generate random patterns on a unit sphere and assess the frequency with which the update rule (8) leads to metastable states of different sizes (Table 1). Figure 1 shows optimization trajectories for several queries and pattern configurations, along with the basins of attraction for the three methods (a larger β is needed to allow the $\alpha = 1$ model to get ϵ-close to a single pattern). We use $\alpha \in \{1, 1.5, 2\}$ since for those cases the Ω_α-RPM admits an exact algorithm (Peters et al., 2019).[2] Additional plots are shown in Appendix D. Overall, we observe that, with $\alpha = 1$ (which corresponds to Ramsauer et al. (2021)) large metastable states abound, whereas for larger α many updates converge exactly to a single pattern.

Table 1: Distribution of metastable state cardinalities (in %), for uniform patterns on a unit sphere ($N = 10, D = 5$), and a uniform query on the unit ball. (Estimated using 1000 random trials; $\beta = 4$; for $\alpha = 1$ we threshold at 0.01).

$\|\boldsymbol{p}\|_0$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$
1	0.0	23.9	72.3
2	0.0	44.4	26.7
3	0.5	25.0	1.0
4	1.9	5.9	0.0
5	9.6	0.8	0.0
6	20.0	0.0	0.0
7	23.5	0.0	0.0
8	25.7	0.0	0.0
9	15.4	0.0	0.0
10	3.4	0.0	0.0

[2] In practice any $\alpha > 1$ can be used with bisection, but we did not want approximations to affect our results.

6 Conclusions and Related Work

We proposed new Hopfield energies, linked to Fenchel-Young losses, which generalize the framework of Ramsauer et al. (2021). For suitable choices of generalized entropy, linked to Tsallis entropies, the resulting update equations mimic the sparse attention mechanism of adaptively sparse transformers (Correia et al., 2019). Our proposed models have interesting properties, including *exact* convergence to single patterns (not just to a nearby region), while maintaining end-to-end differentiability. We provide theoretical conditions for convergence in one iteration, along with toy experiments highlighting the usefulness of the new energies.

Concurrently to our work, Hu et al. (2023) recently proposed a model for sparse modern Hopfield networks along with a memory retrieval error bound provably tighter than the dense analog of Ramsauer et al. (2021). Their energy can be seen as a particular case of our Hopfield-Fenchel-Young energy, specifically the $\alpha = 2$ case (sparsemax). Our work differs in that we consider the more general scenario where $\alpha > 1$ (α-entmax), we make a connection to Fenchel-Young losses, and we use the margin property of α-entmax to prove exact convergence to single patterns under the conditions of Proposition 2. Future work will examine the suitability of the proposed approach to real-world problems, by exploring the use of learnable sparse Hopfield layers in concrete applications.

Acknowledgments

This work was supported by the European Research Council (DECOLLAGE, ERC-2022-CoG 101088763), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020, and by the Dutch Research Council (NWO) via VI.Veni.212.228.

References

Blondel, M., Martins, A. F., and Niculae, V. (2020). Learning with Fenchel-Young losses. *The Journal of Machine Learning Research*, 21(1):1314–1382.

Correia, G. M., Niculae, V., and Martins, A. F. (2019). Adaptively sparse transformers. In *Proceedings of EMNLP-IJCNLP*.

Demircigil, M., Heusel, J., Löwe, M., Upgang, S., and Vermet, F. (2017). On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Hu, J. Y.-C., Yang, D., Wu, D., Xu, C., Chen, B.-Y., and Liu, H. (2023). On sparse modern hopfield model. *arXiv preprint arXiv:2309.12673*.

Krotov, D. and Hopfield, J. J. (2016). Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29.

Leutgeb, J., Leutgeb, S., Moser, M., and Moser, E. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science*, 315(5814):961–966.

Martins, A. and Astudillo, R. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of ICML*.

Neunuebel, J. and Knierim, J. (2014). CA3 retrieves coherent representations from degraded input: direct evidence for CA3 pattern completion and dentate gyrus pattern separation. *Neuron*, 81(2):416–427.

Palm, G. (2013). Neural associative memories and sparse coding. *Neural Netw*, 37:165–171.

Peters, B., Niculae, V., and Martins, A. F. (2019). Sparse sequence-to-sequence models. In *Proceedings of ACL*.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., Klambauer, G., Brandstetter, J., and Hochreiter, S. (2021). Hopfield networks is all you need. In *Proceedings of ICLR*.

Severa, W., Parekh, O., James, C., and Aimone, J. (2017). A combinatorial model for dentate gyrus sparse coding. *Neural Computation*, 29(1):94–117.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216.

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

Tyulmankov, D., Fang, C., Vadaparty, A., and Yang, G. R. (2021). Biological learning in key-value memory networks. *Advances in Neural Information Processing Systems*.

Whittington, J. C., Warren, J., and Behrens, T. E. (2021). Relating transformers to models and neural representations of the hippocampal formation. In *Proceedings of ICLR*.

Yassa, M. and Stark, C. (2011). Pattern separation in the hippocampus. *Trends Neurosci*, 34(10):515–525.

Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural computation*, 15(4):915–936.

A Generalized negentropies

Definition 1. *(Blondel et al., 2020, §4.1). A function $\Omega : \triangle_{N-1} \to \mathbb{R}$ is a generalized negentropy iff it satisfies the following properties:*

1. *Zero negentropy: $\Omega(\boldsymbol{p}) = 0$ if $\boldsymbol{p}$ is a one-hot vector (delta distribution), i.e., $\boldsymbol{p} = \boldsymbol{e}_i$ for any $i \in \{1, \dots, N\}$.*
2. *Strict convexity: $\Omega\left((1-\alpha)\boldsymbol{p} + \alpha\boldsymbol{p}'\right) < (1-\alpha)\Omega(\boldsymbol{p}) + \alpha\Omega(\boldsymbol{p}')$.*
3. *Permutation invariance: $\Omega(\boldsymbol{P}\boldsymbol{p}) = \Omega(\boldsymbol{p})$ for any permutation matrix $\boldsymbol{P}$ (i.e., square matrices with a single 1 in each row and each column, zero elsewhere).*

B Proof of Proposition 1

We start by proving that $E(\boldsymbol{q}) \geq 0$. We show first that for any Ω satisfying conditions 1–3 above, we have

$$L_\Omega(\boldsymbol{\theta}; \mathbf{1}/N) \leq \max_i \theta_i - \mathbf{1}^\top\boldsymbol{\theta}/N. \tag{9}$$

From the definition of Ω^* and the fact that $\Omega(\boldsymbol{p}) \geq \Omega(\mathbf{1}/N)$ for any $\boldsymbol{p} \in \triangle_{N-1}$, we have that, for any $\boldsymbol{\theta}$, $\Omega^*(\boldsymbol{\theta}) = \max_{\boldsymbol{p}\in\triangle_{N-1}} \boldsymbol{\theta}^\top\boldsymbol{p} - \Omega(\boldsymbol{p}) \leq \max_{\boldsymbol{p}\in\triangle_{N-1}} \boldsymbol{\theta}^\top\boldsymbol{p} - \Omega(\mathbf{1}/N) \leq \max_i \theta_i - \Omega(\mathbf{1}/N)$, which leads to (9).

Let now $k = \arg\max_i \boldsymbol{q}^\top\boldsymbol{x}_i$, i.e., $\boldsymbol{x}_k$ is the pattern most similar to the query $\boldsymbol{q}$. Therefore, we have

$$\begin{aligned}
E(\boldsymbol{q}) &= -\beta^{-1}L_\Omega(\beta\boldsymbol{X}\boldsymbol{q}; \mathbf{1}/N) + \frac{1}{2}\|\boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}}\|^2 + \frac{1}{2}(M^2 - \|\boldsymbol{\mu}_{\boldsymbol{X}}\|^2) \\
&\geq -\beta^{-1}(\beta\max_i \boldsymbol{q}^\top\boldsymbol{x}_i - \beta\mathbf{1}^\top\boldsymbol{X}\boldsymbol{q}/N) + \frac{1}{2}\|\boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}}\|^2 + \frac{1}{2}(M^2 - \|\boldsymbol{\mu}_{\boldsymbol{X}}\|^2) \\
&= -\boldsymbol{q}^\top\boldsymbol{x}_k + \boldsymbol{q}^\top\boldsymbol{\mu}_{\boldsymbol{X}} + \frac{1}{2}\|\boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}}\|^2 + \frac{1}{2}(M^2 - \|\boldsymbol{\mu}_{\boldsymbol{X}}\|^2) \\
&= -\boldsymbol{q}^\top\boldsymbol{x}_k + \frac{1}{2}\|\boldsymbol{q}\|^2 + \frac{1}{2}M^2 \\
&\geq -\boldsymbol{q}^\top\boldsymbol{x}_k + \frac{1}{2}\|\boldsymbol{q}\|^2 + \frac{1}{2}\|\boldsymbol{x}_k\|^2 \\
&= \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{q}\|^2 \geq 0.
\end{aligned}$$

The zero value of energy is attained when $\boldsymbol{X} = \mathbf{1}\boldsymbol{q}^\top$ (all patterns are equal to the query), in which case $\boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{q}$, $M = \|\boldsymbol{q}\| = \|\boldsymbol{\mu}_{\boldsymbol{X}}\|$, and we get $E_{\text{convex}}(\boldsymbol{q}) = E_{\text{concave}}(\boldsymbol{q}) = 0$.

Now we prove the two upper bounds. For that, note that, for any $\boldsymbol{p} \in \triangle_{N-1}$, we have $0 \leq L_\Omega(\boldsymbol{\theta}, \boldsymbol{p}) = L_\Omega(\boldsymbol{\theta}, \mathbf{1}/N) - \Omega(\mathbf{1}/N) + \Omega(\boldsymbol{p}) - (\boldsymbol{p} - \mathbf{1}/N)^\top\boldsymbol{\theta} \leq L_\Omega(\boldsymbol{\theta}, \mathbf{1}/N) - \Omega(\mathbf{1}/N) - (\boldsymbol{p} - \mathbf{1}/N)^\top\boldsymbol{\theta}$, due to the assumptions 1–3 which ensure Ω is non-positive. That is, $L_\Omega(\boldsymbol{\theta}, \mathbf{1}/N) \geq \Omega(\mathbf{1}/N) + (\boldsymbol{p} - \mathbf{1}/N)^\top\boldsymbol{\theta}$. Therefore, with $\boldsymbol{q} = \boldsymbol{X}^\top\boldsymbol{p}$, we get

$$E_{\text{concave}}(\boldsymbol{q}) \leq -\beta^{-1}\Omega(\mathbf{1}/N) - \boldsymbol{p}^\top\boldsymbol{X}\boldsymbol{q} + \boldsymbol{q}^\top\boldsymbol{\mu}_{\boldsymbol{X}} = -\beta^{-1}\Omega(\mathbf{1}/N) - \|\boldsymbol{q}\|^2 + \boldsymbol{q}^\top\boldsymbol{\mu}_{\boldsymbol{X}},$$

and $E(\boldsymbol{q}) = E_{\text{concave}}(\boldsymbol{q}) + E_{\text{convex}}(\boldsymbol{q}) \leq -\beta^{-1}\Omega(\mathbf{1}/N) - \|\boldsymbol{q}\|^2 + \boldsymbol{q}^\top\boldsymbol{\mu}_{\boldsymbol{X}} + \frac{1}{2}\|\boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}}\|^2 + \frac{1}{2}(M^2 - \|\boldsymbol{\mu}_{\boldsymbol{X}}\|^2) = -\beta^{-1}\Omega(\mathbf{1}/N) - \frac{1}{2}\|\boldsymbol{q}\|^2 + \frac{1}{2}M^2 \leq -\beta^{-1}\Omega(\mathbf{1}/N) + \frac{1}{2}M^2$.

To show the second upper bound, use the fact that $E_{\text{concave}}(\boldsymbol{q}) \leq 0$, which leads to $E(\boldsymbol{q}) \leq E_{\text{convex}}(\boldsymbol{q}) = \frac{1}{2}\|\boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}}\|^2 + \frac{1}{2}(M^2 - \|\boldsymbol{\mu}_{\boldsymbol{X}}\|^2) = \frac{1}{2}\|\boldsymbol{q}\|^2 - \boldsymbol{q}^\top\boldsymbol{\mu}_{\boldsymbol{X}} + \frac{1}{2}M^2$. Note that $\|\boldsymbol{q}\| = \|\boldsymbol{X}^\top\boldsymbol{p}\| \leq \sum_i p_i\|x_i\| \leq M$ and that, from the Cauchy-Schwarz inequality, we have $-\boldsymbol{q}^\top\boldsymbol{\mu}_{\boldsymbol{X}} \leq \|\boldsymbol{\mu}_{\boldsymbol{X}}\|\|\boldsymbol{q}\| \leq M^2$. Therefore, we obtain $E(\boldsymbol{q}) \leq \frac{1}{2}\|\boldsymbol{q}\|^2 - \boldsymbol{q}^\top\boldsymbol{\mu}_{\boldsymbol{X}} + \frac{1}{2}M^2 \leq \frac{1}{2}M^2 + M^2 + \frac{1}{2}M^2 = 2M^2$.

We now turn to the update rule. The CCCP algorithm works as follows: at the t^{th} iteration, it linearizes the concave function E_{concave} by using a first-order Taylor approximation around $\boldsymbol{q}_t$,

$$E_{\text{concave}}(\boldsymbol{q}) \approx \tilde{E}_{\text{concave}}(\boldsymbol{q}) := E_{\text{concave}}(\boldsymbol{q}_t) + \left(\frac{\partial E_{\text{concave}}(\boldsymbol{q}_t)}{\partial \boldsymbol{q}}\right)^\top (\boldsymbol{q} - \boldsymbol{q}_t).$$

Then, it computes a new iterate by solving the convex optimization problem $\boldsymbol{q}_{t+1} := \arg\min_{\boldsymbol{q}} E_{\text{convex}}(\boldsymbol{q}) + \tilde{E}_{\text{concave}}(\boldsymbol{q})$, which leads to the equation $\nabla E_{\text{convex}}(\boldsymbol{q}_{t+1}) = -\nabla E_{\text{concave}}(\boldsymbol{q}_t)$. Using the fact that $\nabla L_\Omega(\boldsymbol{\theta}, \boldsymbol{p}) = \hat{\boldsymbol{p}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{p}$ and the chain rule leads to

$$\begin{aligned}
\nabla E_{\text{concave}}(\boldsymbol{q}) &= -\beta^{-1} \nabla_{\boldsymbol{q}} L_\Omega(\beta \boldsymbol{X}\boldsymbol{q}; \mathbf{1}/N) = \boldsymbol{X}^\top(\mathbf{1}/N - \hat{\boldsymbol{p}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q})) \\
&= \boldsymbol{\mu}_{\boldsymbol{X}} - \boldsymbol{X}^\top \hat{\boldsymbol{p}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q}) \\
\nabla E_{\text{convex}}(\boldsymbol{q}) &= \boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}},
\end{aligned} \tag{10}$$

giving the update equation (8).

C Proof of Proposition 2

A stationary point is a solution of the equation $-\nabla E_{\text{concave}}(\boldsymbol{q}) = \nabla E_{\text{convex}}(\boldsymbol{q})$. Using the expression for gradients (10), this is equivalent to $\boldsymbol{q} = \boldsymbol{X}^\top \hat{\boldsymbol{p}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q})$. If $\boldsymbol{x}_i = \boldsymbol{X}^\top \boldsymbol{e}_i$ is not a convex combination of the other memory patterns, $\boldsymbol{x}_i$ is a stationary point iff $\hat{\boldsymbol{p}}_\Omega(\beta \boldsymbol{X}\boldsymbol{x}_i) = \boldsymbol{e}_i$. We now use the margin property of α-entmax transformations (6), according to which the latter is equivalent to $\beta \boldsymbol{x}_i^\top \boldsymbol{x}_i - \max_{j \neq i} \beta \boldsymbol{x}_i^\top \boldsymbol{x}_j \geq \frac{1}{\alpha - 1}$. Noting that the left hand side equals $\beta \Delta_i$ leads to the desired result.

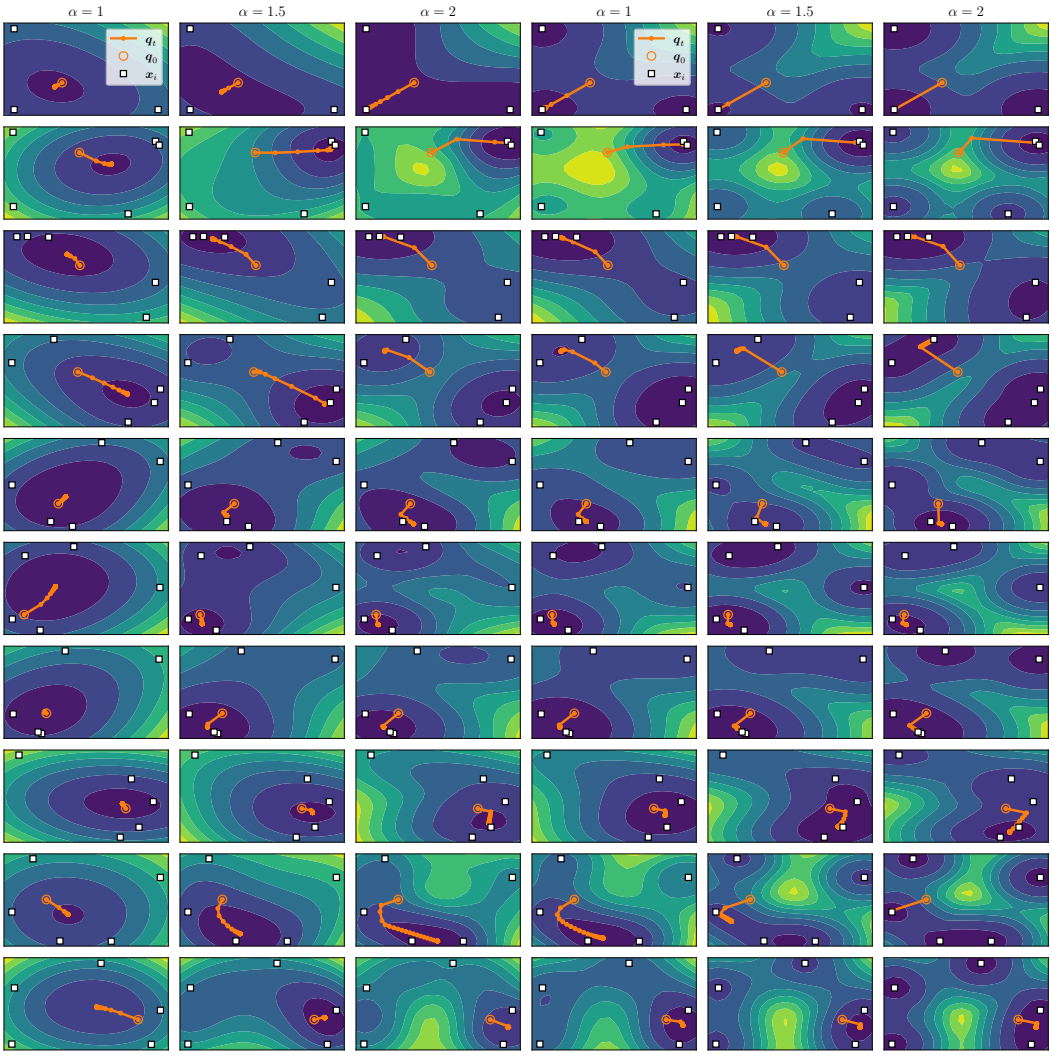
If the initial query satisfies $\boldsymbol{q}_0^\top(\boldsymbol{x}_i - \boldsymbol{x}_{\boldsymbol{j}}) \geq \frac{1}{(\alpha-1)\beta}$ for all $j \neq i$, we have again from the margin property that $\hat{\boldsymbol{p}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q}_0) = \boldsymbol{e}_i$, which combined to the previous claim ensures convergence in one step to $\boldsymbol{x}_i$.

Finally, note that, if $\boldsymbol{q}_0$ is ϵ-close to $\boldsymbol{x}_i$, we have $\boldsymbol{q}_0 = \boldsymbol{x}_i + \epsilon \boldsymbol{r}$ for some vector $\boldsymbol{r}$ with $\|\boldsymbol{r}\| = 1$. Therefore, we have

$$\begin{aligned}
\boldsymbol{q}_0^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) &= (\boldsymbol{x}_i + \epsilon \boldsymbol{r})^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \\
&\geq \Delta_i + \epsilon \boldsymbol{r}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \\
&\geq \Delta_i - \epsilon \underbrace{\|\boldsymbol{r}\|}_{=1} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|,
\end{aligned} \tag{11}$$

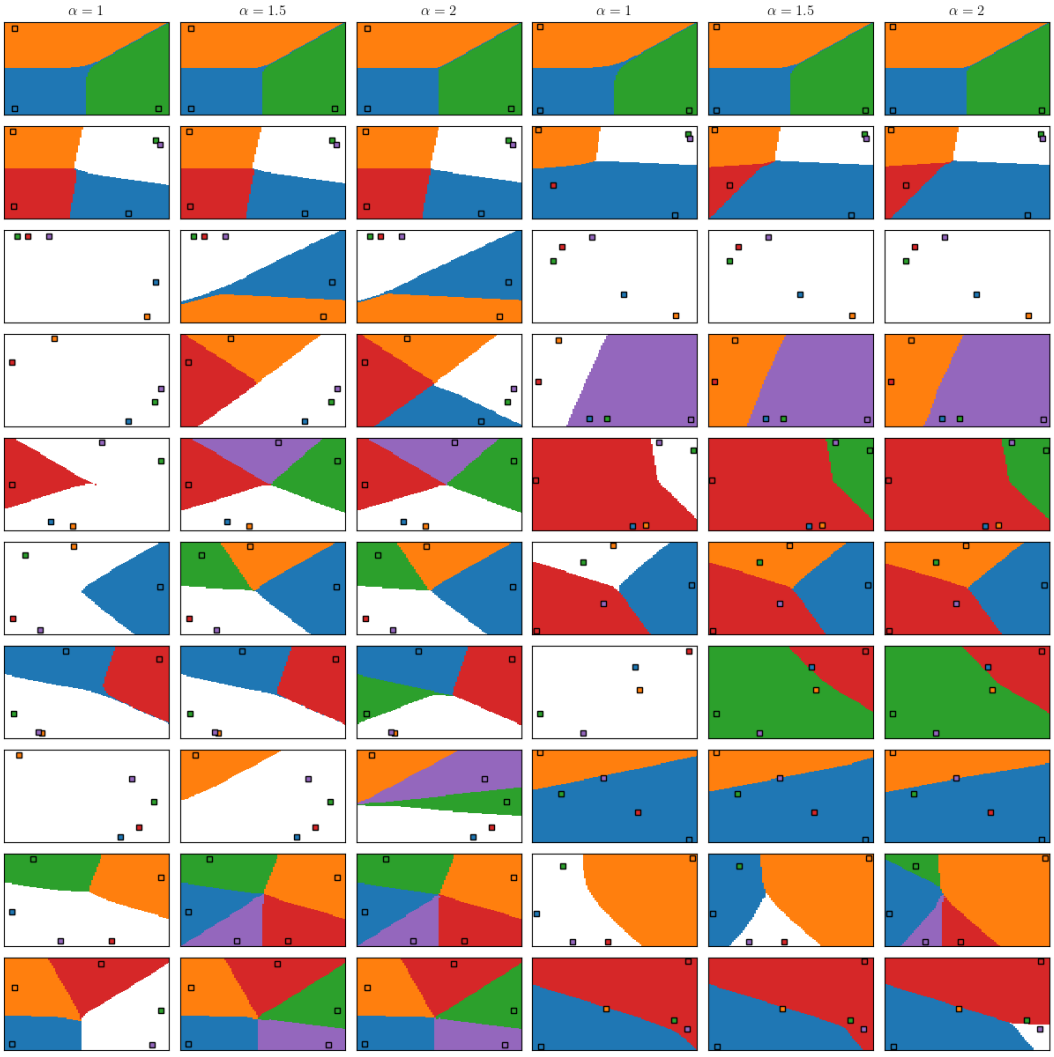
where we invoked the Cauchy-Schwarz inequality in the last step. Since the patterns are normalized, we have from the triangle inequality that $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq \|\boldsymbol{x}_i\| + \|\boldsymbol{x}_j\| = 2$; using the assumption that $\Delta_i \geq \frac{1}{(\alpha-1)\beta} + 2\epsilon$, we obtain $\boldsymbol{q}_0^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq \frac{1}{(\alpha-i)\beta}$, which from the previous points ensures convergence to $\boldsymbol{x}_i$ in one iteration.

D Additional Plots



(a) $\beta = 1$ (b) $\beta = 4$

Figure 2: Additional energy contour and CCCP optimization trajectory plots.



(a) $\beta = 10$, normalized patterns (b) $\beta = 4$, non-normalized patterns

Figure 3: Additional attraction basin plots. A tolerance of $\epsilon = 0.01$ is allowed when $\alpha = 1$.