Sparse and Structured Hopfield Networks

Saul Santos [1 2] **Vlad Niculae** [3] **Daniel McNamee** [4] **André F. T. Martins** [1 2 5]

Abstract

Modern Hopfield networks have enjoyed recent interest due to their connection to attention in transformers. Our paper provides a unified framework for sparse Hopfield networks by establishing a link with Fenchel-Young losses. The result is a new family of Hopfield-Fenchel-Young energies whose update rules are end-to-end differentiable *sparse* transformations. We reveal a connection between loss margins, sparsity, and *exact* memory retrieval. We further extend this framework to *structured* Hopfield networks via the SparseMAP transformation, which can retrieve pattern associations instead of a single pattern. Experiments on multiple instance learning and text rationalization demonstrate the usefulness of our approach.

1. Introduction

Hopfield networks are a kind of biologically-plausible neural network with 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 capacity of classical Hopfield networks was recently overcome through new energy functions and continuous state patterns (Krotov & 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, but their model is incapable of exact retrieval and may require a low temperature to avoid metastable states (states which mix multiple input patterns).

In this paper, we bridge this gap by making a connection between Hopfield energies and **Fenchel-Young losses** (Blondel et al., 2020). We extend the energy functions of Ramsauer et al. (2021) and Hu et al. (2023) to a wider family induced by generalized entropies. The minimization of these energy functions leads to **sparse update rules** which include as particular cases α-entmax (Peters et al., 2019), α-normmax (Blondel et al., 2020), and SparseMAP (Niculae et al., 2018), where a structural constraint ensures the retrieval of k patterns. Unlike Ramsauer et al. (2021)'s Hopfield layers, our update rules pave the way for **exact retrieval**, leading to the emergence of sparse association among patterns while ensuring end-to-end differentiability.

This endeavour aligns with the strong neurobiological motivation to seek new Hopfield energies capable of **sparse** and **structured** retrieval. Indeed, sparse neural activity patterns forming structured representations underpin core principles of cortical computation (Simoncelli & Olshausen, 2001; Tse et al., 2007; Palm, 2013). With respect to memory formation circuits, the sparse firing of neurons in the dentate gyrus (DG), a distinguished region within the hippocampal network, underpins its proposed role in pattern separation during memory storage (Yassa & Stark, 2011; Severa et al., 2017). Evidence suggests that such sparsified activity aids in minimizing interference, however an integrative theoretical account linking sparse coding and attractor network functionality to clarify these empirical observations is lacking (Leutgeb et al., 2007; Neunuebel & Knierim, 2014). Our main contributions are:

- We introduce **Hopfield-Fenchel-Young** energy functions as a generalization of modern Hopfield networks (§3.1).
- We leverage properties of Fenchel-Young losses which relate **sparsity** and **margins** to obtain new theoretical results for exact memory retrieval, proving exponential storage capacity in a stricter sense than prior work (§3.2).
- We propose new **structured** Hopfield networks via the SparseMAP transformation, which return **pattern associations** instead of single patterns. We show that SparseMAP has a structured margin and provide sufficient conditions for exact retrieval of pattern associations (§4).

Experiments on synthetic and real-world tasks (multiple instance learning and text rationalization) showcase the usefulness of our proposed models using various kinds of sparse

[*]Equal contribution [1]Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal [2]Instituto de Telecomunicações, Lisbon, Portugal [3]Language Technology Lab, University of Amsterdam, The Netherlands [4]Champalimaud Research, Lisbon, Portugal [5]Unbabel, Lisbon, Portugal. Correspondence to: Saul Santos <saul.r.santos@tecnico.ulisboa.pt>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

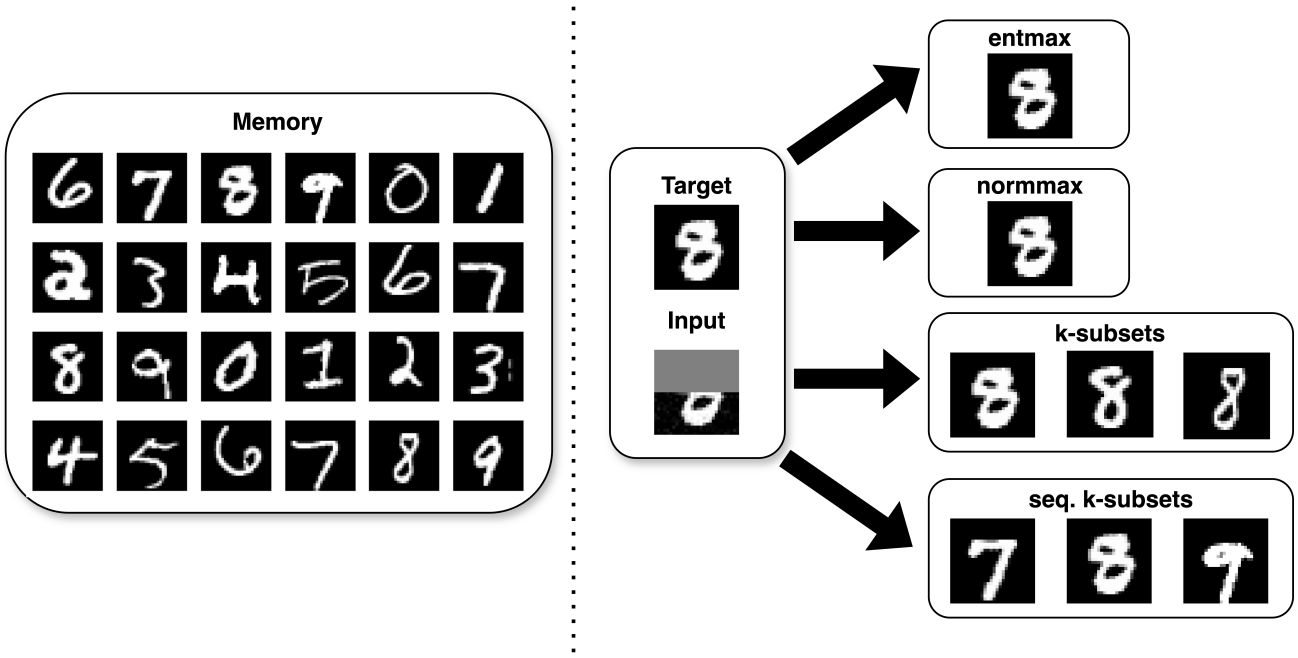

Figure 1: Overview of the Hopfield networks proposed in this paper: sparse transformations (entmax and normmax) aim to retrieve the closest pattern to the query, and they have exact retrieval guarantees. Structured variants find pattern associations. The k-subsets transformation returns a mixture of the top-k patterns, and sequential k-subsets favors contiguous retrieval.

and structured transformations (§5). An overview of the utility of our methods can be seen in Fig. 1.[1]

Notation. We denote by $\triangle_N$ the $(N-1)^{\text{th}}$-dimensional probability simplex, $\triangle_N := \{\boldsymbol{p} \in \mathbb{R}^N : \boldsymbol{p} \geq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{p} = 1\}$. The convex hull of a set $\mathcal{Y} \subseteq \mathbb{R}^M$ is $\operatorname{conv}(\mathcal{Y}) := \{\sum_{i=1}^N p_i \boldsymbol{y}_i : \boldsymbol{p} \in \triangle_N, \boldsymbol{y}_1, ..., \boldsymbol{y}_N \in \mathcal{Y}, N \in \mathbb{N}\}$. We have $\triangle_N = \operatorname{conv}(\{\boldsymbol{e}_1, ..., \boldsymbol{e}_N\})$, where $\boldsymbol{e}_i \in \mathbb{R}^N$ is the ith basis (one-hot) vector. Given a convex function $\Omega : \mathbb{R}^N \to \bar{\mathbb{R}}$, where $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$, we denote its domain by $\operatorname{dom}(\Omega) := \{\boldsymbol{y} \in \mathbb{R}^N : \Omega(\boldsymbol{y}) < +\infty\}$ and its Fenchel conjugate by $\Omega^*(\boldsymbol{\theta}) = \sup_{\boldsymbol{y} \in \mathbb{R}^N} \boldsymbol{y}^\top \boldsymbol{\theta} - \Omega(\boldsymbol{y})$. We denote by $I_{\mathcal{C}}$ the indicator function of a convex set $\mathcal{C}$, defined as $I_{\mathcal{C}}(\boldsymbol{y}) = 0$ if $\boldsymbol{y} \in \mathcal{C}$, and $I_{\mathcal{C}}(\boldsymbol{y}) = +\infty$ otherwise.

2. Background

2.1. Hopfield Networks

Let $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ be a matrix whose rows hold a set of examples $\boldsymbol{x}_1, ..., \boldsymbol{x}_N \in \mathbb{R}^D$ ("memory patterns") and let $\boldsymbol{q}^{(0)} \in \mathbb{R}^D$ be a query vector (or "state pattern"). Hopfield networks iteratively update $\boldsymbol{q}^{(t)} \mapsto \boldsymbol{q}^{(t+1)}$ for $t \in \{0, 1, ...\}$ according to a certain rule, eventually converging 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 $E(\boldsymbol{q}) = -\frac{1}{2}\boldsymbol{q}^\top \boldsymbol{W}\boldsymbol{q}$, with $\boldsymbol{q} \in \{\pm 1\}^D$ and $\boldsymbol{W} = \boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{D \times D}$, leading to the update rule $\boldsymbol{q}^{(t+1)} = \operatorname{sign}(\boldsymbol{W}\boldsymbol{q}^{(t)})$. A limitation of this classical network is that it has only $N = \mathcal{O}(D)$ memory storage capacity, above which patterns start to interfere (Amit et al., 1985; McEliece et al., 1987).

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

$$E(\boldsymbol{q}) = -\frac{1}{\beta} \log \sum_{i=1}^N \exp(\beta \boldsymbol{x}_i^\top \boldsymbol{q}) + \frac{1}{2}\|\boldsymbol{q}\|^2 + \text{const.} \quad (1)$$

Ramsauer et al. (2021) 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 & Rangarajan 2003) leads to the update rule:

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{X}^\top \operatorname{softmax}(\beta \boldsymbol{X}\boldsymbol{q}^{(t)}). \quad (2)$$

When $\beta = \frac{1}{\sqrt{D}}$, each update matches the computation per-

[1] Our code is available on https://github.com/deep-spin/SSHN

formed 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 (Hoover et al., 2023).

While Ramsauer et al. (2021) have derived useful theoretical properties of these networks (including their exponential storage capacity under a relaxed notion of retrieval), the use of softmax in (2) prevents exact convergence and may lead to undesirable metastable states. Recently (and closely related to our work), Hu et al. (2023) introduced sparse Hopfield networks and proved favorable retrieval properties, but still under an approximate sense, where attractors are close but distinct from the stored patterns. We overcome these drawbacks in our paper, where we provide a unified treatment of sparse and structured Hopfield networks along with theoretical analysis showing that exact retrieval is possible without sacrificing exponential storage capacity.

2.2. Sparse Transformations and Fenchel-Young Losses

Our construction and results flow from the properties of Fenchel-Young losses, which we next review.

Given a strictly convex function $\Omega : \mathbb{R}^N \to \bar{\mathbb{R}}$ with domain $\mathrm{dom}(\Omega)$, the Ω-regularized argmax transformation (Niculae & Blondel, 2017), $\hat{\boldsymbol{y}}_\Omega : \mathbb{R}^N \to \mathrm{dom}(\Omega)$, is:

$$\hat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) := \nabla\Omega^*(\boldsymbol{\theta}) = \underset{\boldsymbol{y}\in\mathrm{dom}(\Omega)}{\mathrm{argmax}}\ \boldsymbol{\theta}^\top\boldsymbol{y} - \Omega(\boldsymbol{y}). \quad (3)$$

Let us assume for now that $\mathrm{dom}(\Omega) = \triangle_N$ (the probability simplex). One instance of (3) is the **softmax** transformation, obtained when the regularizer is the Shannon negentropy, $\Omega(\boldsymbol{y}) = \sum_{i=1}^N y_i \log y_i + I_{\triangle_N}(\boldsymbol{y})$. Another instance is the **sparsemax** transformation, obtained when $\Omega(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2 + I_{\triangle_N}(\boldsymbol{y})$ (Martins & Astudillo, 2016). The sparsemax transformation is 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 corresponds to the following choice of regularizer, called the **Tsallis α-negentropy** (Tsallis, 1988):

$$\Omega_\alpha^T(\boldsymbol{y}) = \frac{-1 + \|\boldsymbol{y}\|_\alpha^\alpha}{\alpha(\alpha-1)} + I_{\triangle_N}(\boldsymbol{y}). \quad (4)$$

When $\alpha \to 1$, the Tsallis α-negentropy Ω_α^T becomes Shannon's negentropy, leading to the softmax transformation, and when $\alpha = 2$, it becomes the ℓ_2-norm (up to a constant) and we recover the sparsemax. Another example is the **norm α-negentropy** (Blondel et al., 2020, §4.3),

$$\Omega_\alpha^N(\boldsymbol{y}) = -1 + \|\boldsymbol{y}\|_\alpha + I_{\triangle_N}(\boldsymbol{y}), \quad (5)$$

which, when $\alpha \to +\infty$, is called the Berger-Parker dominance index (May, 1975). We call the resulting transformation **α-normmax**. While the Tsallis and norm negentropies

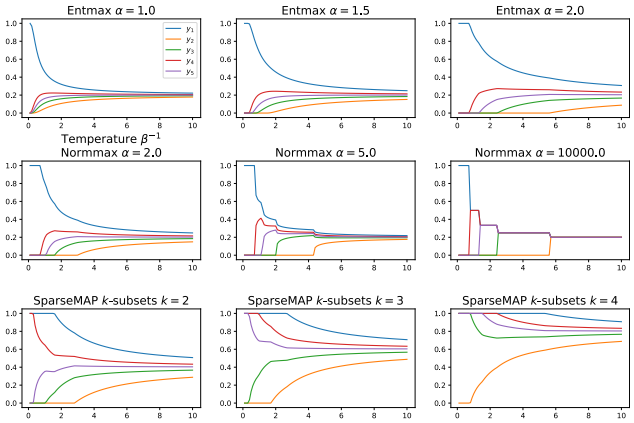

Figure 2: Sparse and structured transformations used in this paper and their regularization path. In each plot, we show $\hat{\boldsymbol{y}}_\Omega(\beta\boldsymbol{\theta}) = \hat{\boldsymbol{y}}_{\beta^{-1}\Omega}(\boldsymbol{\theta})$ as a function of the temperature β^{-1} where $\boldsymbol{\theta} = [1.0716, -1.1221, -0.3288, 0.3368, 0.0425]^\top$. Additional examples can be found in App. E.1.

have similar expressions and the resulting transformations both tend to be sparse, they have important differences, as suggested in Fig. 2: normmax favors distributions closer to uniform over the selected support.

The **Ω-Fenchel-Young loss** (Blondel et al., 2020) is

$$L_\Omega(\boldsymbol{\theta}, \boldsymbol{y}) := \Omega(\boldsymbol{y}) + \Omega^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top\boldsymbol{y}. \quad (6)$$

When Ω is Shannon's negentropy, we have $\Omega^*(\boldsymbol{\theta}) = \log\sum_{i=1}^N \exp(\theta_i)$, and L_Ω is the **cross-entropy loss**, up to a constant. Intuitively, Fenchel-Young losses quantify how "compatible" a score vector $\boldsymbol{\theta} \in \mathbb{R}^N$ (e.g., logits) is to a desired target $\boldsymbol{y} \in \mathrm{dom}(\Omega)$ (e.g., a probability vector).

Fenchel-Young losses have relevant properties (Blondel et al., 2020, Prop. 2): (i) they are non-negative, $L_\Omega(\boldsymbol{\theta}, \boldsymbol{y}) \geq 0$, with equality iff $\boldsymbol{y} = \hat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$; (ii) they are convex on $\boldsymbol{\theta}$ and their gradient is $\nabla_{\boldsymbol{\theta}} L_\Omega(\boldsymbol{\theta}, \boldsymbol{y}) = -\boldsymbol{y} + \hat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$. But, most importantly, they have well-studied **margin** properties, which, as we shall see, will play a pivotal role in this work.

Definition 1 (Margin)**.** *A loss function $L(\boldsymbol{\theta}; \boldsymbol{y})$ has a **margin** if there exists a finite $m \geq 0$ such that*

$$\forall i \in [N], \quad L(\boldsymbol{\theta}, \boldsymbol{e}_i) = 0 \iff \theta_i - \max_{j\neq i}\theta_j \geq m. \quad (7)$$

The smallest such m is called the margin of L. If L_Ω is a Fenchel-Young loss, (7) is equivalent to $\hat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \boldsymbol{e}_i$.

A famous example of a loss with a margin of 1 is the hinge loss of support vector machines. On the other hand, the cross-entropy loss does not have a margin. Blondel et al. (2020, Prop. 7) have shown that Tsallis negentropies Ω_α^T with $\alpha > 1$ have a margin of $m = (\alpha-1)^{-1}$, and that norm-entropies Ω_α^N with $\alpha > 1$ have a margin $m = 1$,

independently of α. We will use these facts in the sequel.

3. Sparse Hopfield-Fenchel-Young Energies

We now use Fenchel-Young losses (6) to define a new class of energy functions for modern Hopfield networks.

3.1. Definition and update rule

We start by assuming that the regularizer Ω has domain $\mathrm{dom}(\Omega) = \triangle_N$ and that it is a **generalized negentropy**, i.e., null when $\boldsymbol{y}$ is a one-hot vector, strictly convex, and permutation-invariant (see Appendix A for details). These conditions imply that $\Omega \leq 0$ and that $\Omega(\boldsymbol{y})$ is minimized when $\boldsymbol{y} = \mathbf{1}/N$ is the uniform distribution (Blondel et al., 2020, Prop. 4). Tsallis negentropies (4) for $\alpha \geq 1$ and norm negentropies (5) for $\alpha > 1$ both satisfy these properties.

We define the **Hopfield-Fenchel-Young (HFY) energy** as

$$E(\boldsymbol{q}) = \underbrace{-\beta^{-1} L_\Omega(\beta \boldsymbol{X}\boldsymbol{q}; \mathbf{1}/N)}_{E_{\text{concave}}(\boldsymbol{q})} + \underbrace{\frac{1}{2}\|\boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}}\|^2 + \frac{1}{2}(M^2 - \|\boldsymbol{\mu}_{\boldsymbol{X}}\|^2)}_{E_{\text{convex}}(\boldsymbol{q})}, \tag{8}$$

where $\boldsymbol{\mu}_{\boldsymbol{X}} := \boldsymbol{X}^\top \mathbf{1}/N \in \mathbb{R}^D$ is the empirical mean of the patterns, and $M := \max_i \|\boldsymbol{x}_i\|$. This energy extends that of (1), which is recovered when Ω is Shannon's negentropy. 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. E_{convex} is convex because it is quadratic.[2]

These two terms compete when minimizing the energy (8):

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

The next result, proved in App. D.1, establishes bounds and derives the Hopfield update rule for energy (8), generalizing Ramsauer et al. (2021, Lemma A.1, Theorem A.1).

Proposition 1 (Update rule of HFY energies)**.** *Let the query $\boldsymbol{q}$ be in the convex hull of the rows of $\boldsymbol{X}$, i.e., $\boldsymbol{q} = \boldsymbol{X}^\top \boldsymbol{y}$ for some $\boldsymbol{y} \in \triangle_N$. Then, the energy (8) 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 (8) with the CCCP algorithm (Yuille & Rangarajan, 2003) leads to the updates:*

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{X}^\top \hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q}^{(t)}). \tag{9}$$

In particular, when $\Omega = \Omega_\alpha^T$ (the Tsallis α-negentropy (4)), the update (9) corresponds to the adaptively sparse transformer of Correia et al. (2019). The α-entmax transformation can be computed in linear time for $\alpha \in \{1, 1.5, 2\}$ and for other values of α an efficient bisection algorithm was proposed by Peters et al. (2019). The case $\alpha = 2$ (sparsemax) corresponds to the sparse modern Hopfield network recently proposed by Hu et al. (2023).

When $\Omega = \Omega_\alpha^N$ (the norm α-negentropy (5)), we obtain the α-normmax transformation. This transformation is harder to compute since Ω_α^N is not separable, but we derive in App. B an efficient bisection algorithm which works for any $\alpha > 1$.

3.2. Margins, sparsity, and exact retrieval

Prior work on modern Hopfield networks (Ramsauer et al., 2021, Def. 1) defines pattern storage and retrieval in an *approximate* sense: they assume a small neighbourhood around each pattern $\boldsymbol{x}_i$ containing an attractor $\boldsymbol{x}_i^*$, such that if the initial query $\boldsymbol{q}^{(0)}$ is close enough, the Hopfield updates will converge to $\boldsymbol{x}_i^*$, leading to a retrieval error of $\|\boldsymbol{x}_i^* - \boldsymbol{x}_i\|$. For this error to be small, a large β may be necessary.

We consider here a stronger definition of **exact retrieval**, where the attractors *coincide* with the actual patterns (rather than being nearby). Our main result is that **zero retrieval error** is possible in HFY networks as long as the corresponding Fenchel-Young loss has a **margin** (Def. 1). Given that $\hat{\boldsymbol{y}}_\Omega$ being a sparse transformation is a sufficient condition for L_Ω having a margin (Blondel et al., 2020, Prop. 6), this is a general statement about sparse transformations.[3]

Definition 2 (Exact retrieval)**.** *A pattern $\boldsymbol{x}_i$ is **exactly retrieved** for query $\boldsymbol{q}^{(0)}$ iff there is a finite number of steps T such that iterating (9) leads to $\boldsymbol{q}^{(T')} = \boldsymbol{x}_i\ \forall T' \geq T$.*

The following result gives sufficient conditions for exact retrieval with $T = 1$ given that patterns are well separated and that the query is sufficiently close to the retrieved pattern. It establishes the exact **autoassociative** property of HFY networks: if all patterns are slightly perturbed, the Hopfield dynamics are able to recover the original patterns exactly. 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 (Exact retrieval in a single iteration)**.** *Assume L_Ω has margin m, and let $\boldsymbol{x}_i$ be a pattern outside the convex hull of the other patterns. Then, $\boldsymbol{x}_i$ is a station-*

[2] Up to constants, for this choice of Ω our convex-concave decomposition is the same as Ramsauer et al. (2021).

[3] At first sight, this might seem a surprising result, given that both queries and patterns are continuous. The reason why exact convergence is possible hinges crucially on sparsity.

ary point of the energy (8) iff $\Delta_i \geq m\beta^{-1}$. In addition, if the initial query $\boldsymbol{q}^{(0)}$ satisfies ${\boldsymbol{q}^{(0)}}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq m\beta^{-1}$ for all $j \neq i$, then the update rule (9) converges to $\boldsymbol{x}_i$ exactly in one iteration. Moreover, if the patterns are normalized, $\|\boldsymbol{x}_i\| = M$ for all i, and well-separated with $\Delta_i \geq m\beta^{-1} + 2M\epsilon$, 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.

The proof is in Appendix D.2. For the Tsallis negentropy case $\Omega = \Omega_\alpha^T$ with $\alpha > 1$ (the sparse case), we have $m = (\alpha-1)^{-1}$ (cf. Def. 1), leading to the condition $\Delta_i \geq \frac{1}{(\alpha-1)\beta}$. 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 $\epsilon = \mathcal{O}(\exp(-\beta))$ requires a low temperature (large β). It is also stronger than the retrieval error bound recently derived by Hu et al. (2023, Theorem 2.2) for the case $\alpha = 2$, which has an additive term involving M and therefore does not provide conditions for exact retrieval.

For the normmax negentropy case $\Omega = \Omega_\alpha^N$ with $\alpha > 1$, we have $m = 1$, so the condition above becomes $\Delta_i \geq \frac{1}{\beta}$.

Given that exact retrieval is a stricter definition, one may wonder whether requiring it sacrifices storage capacity. Reassuringly, the next result, inspired but stronger than Ramsauer et al. (2021, Theorem A.3), shows that HFY networks with exact retrieval also have exponential storage capacity.

Proposition 3 (Storage capacity with exact retrieval)**.** *Assume patterns are placed equidistantly on the sphere of radius M. The HFY network can store and exactly retrieve $N = \mathcal{O}((2/\sqrt{3})^D)$ patterns in one iteration under a ϵ-perturbation as long as $M^2 > 2m\beta^{-1}$ and*

$$\epsilon \leq \frac{M}{4} - \frac{m}{2\beta M}. \tag{10}$$

Assume patterns are randomly placed on the sphere with uniform distribution. Then, with probability $1-p$, the HFY network can store and exactly retrieve $N = \mathcal{O}(\sqrt{p}\zeta^{\frac{D-1}{2}})$ patterns in one iteration under a ϵ-perturbation if

$$\epsilon \leq \frac{M}{2}\left(1 - \cos\frac{1}{\zeta}\right) - \frac{m}{2\beta M}. \tag{11}$$

The proof is in Appendix D.3.

4. Structured Hopfield Networks

In §3, we considered the case where $\boldsymbol{y} \in \mathrm{dom}(\Omega) = \triangle_N$, the scenario studied by Ramsauer et al. (2021) and Hu et al. (2023). We now take one step further and consider the more general **structured** case, where $\boldsymbol{y}$ is a vector of "marginals" associated to some given structured set. This structure can reflect **pattern associations** that we might want to induce when querying the Hopfield network with $\boldsymbol{q}^{(0)}$. Possible structures include k**-subsets** of memory patterns, potentially leveraging **sequential memory structure**, tree structures, matchings, etc. In these cases, the set of pattern associations we can form is combinatorial, hence it can be considerably larger than the number N of memory patterns.

4.1. Unary scores and structured constraints

We consider first a simpler scenario where there is a predefined set of structures $\mathcal{Y} \subseteq \{0,1\}^N$ and N unary scores, one for each memory pattern. We show in §4.2 how this framework can be generalized for higher-order interactions modeling soft interactions among patterns.

Let $\mathcal{Y} \subseteq \{0,1\}^N$ be a set of binary vectors indicating the underlying set of structures, and let $\mathrm{conv}(\mathcal{Y}) \subseteq [0,1]^N$ denote its convex hull, called the **marginal polytope** associated to the structured set $\mathcal{Y}$ (Wainwright et al., 2008).

Example 1 (k-subsets)**.** *We may be interested in retrieving subsets of k patterns, e.g., to take into account a k-ary relation among patterns or to perform top-k retrieval. In this case, we define $\mathcal{Y} = \{\boldsymbol{y} \in \{0,1\}^N : \boldsymbol{1}^\top\boldsymbol{y} = k\}$, where $k \in [N]$. If $k = 1$, we get $\mathcal{Y} = \{\boldsymbol{e}_1, ..., \boldsymbol{e}_N\}$ and $\mathrm{conv}(\mathcal{Y}) = \triangle_N$, recovering the scenario studied in §3. For larger k, $|\mathcal{Y}| = \binom{N}{k} \gg N$. With a simple rescaling, the resulting marginal polytope is equivalent to the capped probability simplex described by Blondel et al. (2020, §7.3).*

Given unary scores $\boldsymbol{\theta} \in \mathbb{R}^N$, the structure with the largest score is $\boldsymbol{y}^* = \mathrm{argmax}_{\boldsymbol{y}\in\mathcal{Y}} \boldsymbol{\theta}^\top\boldsymbol{y} = \mathrm{argmax}_{\boldsymbol{y}\in\mathrm{conv}(\mathcal{Y})} \boldsymbol{\theta}^\top\boldsymbol{y}$, where the last equality comes from the fact that $\mathrm{conv}(\mathcal{Y})$ is a polytope, therefore the maximum is attained at a vertex. As in (3), we consider a regularized prediction version of this problem via a convex regularizer $\Omega : \mathrm{conv}(\mathcal{Y}) \to \mathbb{R}$:

$$\hat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) := \underset{\boldsymbol{y}\in\mathrm{conv}(\mathcal{Y})}{\mathrm{argmax}}\ \boldsymbol{\theta}^\top\boldsymbol{y} - \Omega(\boldsymbol{y}). \tag{12}$$

By choosing $\Omega(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2 + I_{\mathrm{conv}(\mathcal{Y})}(\boldsymbol{y})$, we obtain **SparseMAP**, a structured version of sparsemax which can be computed efficiently via an active set algorithm as long as an argmax oracle is available (Niculae et al., 2018).

4.2. General case: factor graph, high order interactions

In general, we may want to consider soft interactions among patterns, for example due to temporal dependencies, hierarchical structure, etc. These interactions can be expressed as a bipartite **factor graph** (V, F), where $V = \{1, ..., N\}$ are variable nodes (associated with the patterns) and $F \subseteq 2^V$ are factor nodes representing the interactions (Kschischang et al., 2001). A structure can be represented as a bit vector $\boldsymbol{y} = [\boldsymbol{y}_V; \boldsymbol{y}_F]$, where $\boldsymbol{y}_V$ and $\boldsymbol{y}_F$ indicate configurations of variable and factor nodes, respectively. The SparseMAP transformation has the same form (12) with

$\Omega(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}_V\|^2 + I_{\mathrm{conv}(\mathcal{Y})}(\boldsymbol{y})$ (note that only the unary variables are regularized). Full details are given in App. C.

Example 2 (sequential k-subsets)**.** *Consider the k-subset problem of Example 1 but now with a sequential structure. This can be represented as a pairwise factor graph (V, F) where $V = \{1, ..., N\}$ and $F = \{(i, i+1)\}_{i=1}^{N-1}$. The budget constraint forces exactly k of the N variable nodes to have the value 1. The set $\mathcal{Y}$ contains all bit vectors satisfying these constraints as well as consistency among the variable and factor assignments. To promote consecutive memory items to be both retrieved or neither retrieved, one can define attractive Ising higher-order (pairwise) scores $\boldsymbol{\theta}_{(i,i+1)}$, in addition to the unary scores. The MAP inference problem can be solved with dynamic programming in runtime $\mathcal{O}(Nk)$, and the SparseMAP transformation can be computed with the active set algorithm (Niculae et al., 2018) by iteratively calling this MAP oracle.*

4.3. Structured Fenchel-Young losses and margins

The notion of margin in Def. 1 can be extended to the structured case (Blondel et al., 2020, Def. 5):

Definition 3 (Structured margin)**.** *A loss $L(\boldsymbol{\theta}; \boldsymbol{y})$ has a **structured margin** if $\exists 0 \leq m < \infty$ such that $\forall \boldsymbol{y} \in \mathcal{Y}$:*

$$\boldsymbol{\theta}^\top \boldsymbol{y} \geq \max_{\boldsymbol{y}' \in \mathcal{Y}} \left(\boldsymbol{\theta}^\top \boldsymbol{y}' + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|^2 \right) \;\Rightarrow\; L(\boldsymbol{\theta}; \boldsymbol{y}) = 0.$$

The smallest such m is called the margin of L.

Note that this generalizes the notion of margin in Def. 1, recovered when $\mathcal{Y} = \{\boldsymbol{e}_1, ..., \boldsymbol{e}_N\}$. Note also that, since we are assuming $\mathcal{Y} \subseteq \{0, 1\}^L$, the term $\|\boldsymbol{y} - \boldsymbol{y}'\|^2$ is a Hamming distance, which counts how many bits need to be flipped to transform $\boldsymbol{y}'$ into $\boldsymbol{y}$. A well-known example of a loss with a structured separation margin is the structured hinge loss (Taskar et al., 2003; Tsochantaridis et al., 2005).

We show below that the SparseMAP loss has a structured margin (our result, proved in App. D.4, extends that of Blondel et al. (2020), who have shown this only for structures without high order interactions):

Proposition 4. *Let $\mathcal{Y} \subseteq \{0, 1\}^L$ be contained in a sphere, i.e., for some $r > 0$, $\|\boldsymbol{y}\| = r$ for all $\boldsymbol{y} \in \mathcal{Y}$. Then:*

1. *If there are no high order interactions, then the SparseMAP loss has a structured margin $m = 1$.*
2. *If there are high order interactions and, for some r_V and r_F with $r_V^2 + r_F^2 = r^2$, we have $\|\boldsymbol{y}_V\| = r_V$ and $\|\boldsymbol{y}_F\| = r_F$ for any $\boldsymbol{y} = [\boldsymbol{y}_V; \boldsymbol{y}_F] \in \mathcal{Y}$, then the SparseMAP loss has a structured margin $m \leq 1$.*

The assumptions above are automatically satisfied with the factor graph construction in §4.2, with $r_V^2 = |V|$, $r_F^2 = |F|$, and $r^2 = |V| + |F|$. For the k-subsets example, we have $r^2 = k$, and for the sequential k-subsets example, we have $r_V^2 = N$, $r_F^2 = N - 1$, and $r^2 = 2N - 1$.

4.4. Guarantees for retrieval of pattern associations

We now consider a structured HFY network using SparseMAP. We obtain the following updates:

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{X}^\top \mathrm{SparseMAP}(\beta \boldsymbol{X} \boldsymbol{q}^{(t)}). \tag{13}$$

In the structured case, we aim to retrieve not individual patterns but pattern associations of the form $\boldsymbol{X}^\top \boldsymbol{y}$, where $\boldsymbol{y} \in \mathcal{Y}$. Naturally, when $\mathcal{Y} = \{\boldsymbol{e}_1, ..., \boldsymbol{e}_N\}$, we recover the usual patterns, since $\boldsymbol{x}_i = \boldsymbol{X}^\top \boldsymbol{e}_i$. We define the separation of pattern association $\boldsymbol{y}_i \in \mathcal{Y}$ from data as $\Delta_i = \boldsymbol{y}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y}_i - \max_{j \neq i} \boldsymbol{y}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y}_j$. The next proposition, proved in App. D.5, states conditions for exact convergence in a single iteration, generalizing Prop. 2.

Proposition 5 (Exact structured retrieval)**.** *Let $\Omega(\boldsymbol{y})$ be the SparseMAP regularizer and assume the conditions of Prop. 4 hold. Let $\boldsymbol{y}_i \in \mathcal{Y}$ be such that $\Delta_i \geq \frac{D_i^2}{2\beta}$, where $D_i = \max \|\boldsymbol{y}_i - \boldsymbol{y}_j\| \leq 2r$. Then, $\boldsymbol{X}^\top \boldsymbol{y}_i$ is a stationary point of the Hopfield energy. In addition, if $\boldsymbol{q}^{(0)}$ satisfies ${\boldsymbol{q}^{(0)}}^\top \boldsymbol{X}^\top (\boldsymbol{y}_i - \boldsymbol{y}_j) \geq \frac{D_i^2}{2\beta}$ for all $j \neq i$, then the update rule (13) converges to the pattern association $\boldsymbol{X}^\top \boldsymbol{y}_i$ in one iteration. Moreover, if*

$$\Delta_i \geq \frac{D_i^2}{2\beta} + \epsilon \min\{\sigma_{\max}(\boldsymbol{X}) D_i, M D_i^2\},$$

where $\sigma_{\max}(\boldsymbol{X})$ is the spectral norm of $\boldsymbol{X}$ and $M = \max_k \|\boldsymbol{x}_k\|$, then any $\boldsymbol{q}^{(0)}$ ϵ-close to $\boldsymbol{X}^\top \boldsymbol{y}_i$ will converge to $\boldsymbol{X}^\top \boldsymbol{y}_i$ in one iteration.

Note that the bound above on Δ_i includes as a particular case the unstructured bound in Prop. 2 applied to sparsemax (entmax with $\alpha = 2$, which has margin $m = 1/(\alpha - 1) = 1$), since for $\mathcal{Y} = \triangle_N$ we have $r = 1$ and $D_i = \sqrt{2}$, which leads to the condition $\Delta_i \geq \beta^{-1} + 2M\epsilon$.

For the particular case of the k-subsets problem (Example 1), we have $r = \sqrt{k}$ and $D_i = \sqrt{2k}$, leading to the condition $\Delta_i \geq \frac{k}{\beta} + 2Mk\epsilon$. This recovers sparsemax when $k = 1$.

For the sequential k-subsets problem in Example 2, we have $r = 2N - 1$. Noting that any two distinct $\boldsymbol{y}$ and $\boldsymbol{y}'$ differ in at most $2k$ variable nodes, and since each variable node can affect 6 bits (2 for $\boldsymbol{y}_V$ and 4 for $\boldsymbol{y}_F$), the Hamming distance between $\boldsymbol{y}$ and $\boldsymbol{y}'$ is at most $12k$, therefore we have $D_i = \sqrt{12k}$, which leads to the condition $\Delta_i \geq \frac{6k}{\beta} + 12Mk\epsilon$.

5. Experiments

We now present experiments using synthetic and real-world datasets to validate our theoretical findings and illustrate the usefulness of sparse and structured Hopfield networks.

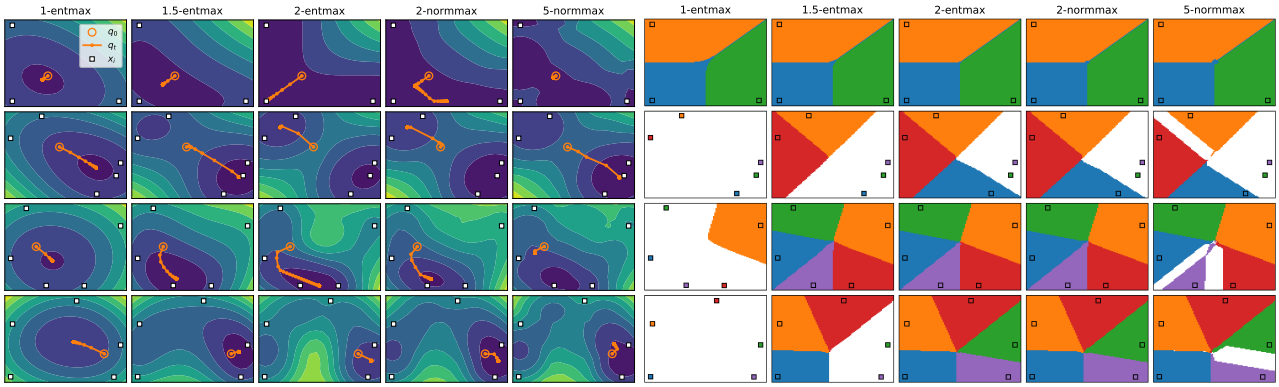

Figure 3: Left: contours of the energy function and optimization trajectory of the CCCP iteration ($\beta = 1$). Right: attraction basins associated with each pattern. (White sections do not converge to a single pattern but to a metastable state; $\beta = 10$ (a larger β is needed to allow for the 1-entmax to get ϵ-close to a single pattern); for $\alpha = 1$ we allow a tolerance of $\epsilon = .01$). Additional plots for different β can be found in App. E.2.

5.1. Hopfield dynamics and basins of attraction

Fig. 3 shows optimization trajectories and basins of attraction for various queries and artificially generated pattern configurations for the two families of sparse transformations, α-entmax and α-normmax. We use $\alpha \in \{1, 1.5, 2\}$ for α-entmax and $\alpha \in \{2, 5\}$ for α-normmax (where we apply the bisection algorithm described in App. B). As α increases, α-entmax converges more often to a single pattern, whereas α-normmax tends to converge towards an attractor which is a uniform average of some patterns.

5.2. Metastable state distributions in MNIST

We next investigate how often our Hopfield networks converge to metastable states, a crucial aspect for understanding the network's dynamics. To elucidate this, we examine $\hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X} \boldsymbol{q}^{(t)})$ for the MNIST dataset, probing the number of nonzeros of these vectors. We set a threshold > 0.01 for the softmax method (1-entmax). For the sparse transformations we do not need a threshold, since they have exact retrieval.

Results in Tab. 1 suggest that α-entmax is capable of retrieving single patterns for higher values of α. Despite α-normmax's ability to induce sparsity, we observe that it tends to stabilize in small but persistent metastable states as α is increased, whereas SparseMAP with k-subsets is capable of retrieving associations of k patterns, as expected.

5.3. Multiple instance learning

In multiple instance learning (MIL), instances are grouped into "bags" and a bag is labeled as positive if it contains at least one instance from a given class. We also consider a extended variant, denoted K-MIL, where bags are considered positive if they contain K or more positive instances.

Ramsauer et al. (2021) tackle MIL via a Hopfield pooling layer, where the query $\boldsymbol{q}$ is learned and the keys $\boldsymbol{X}$ are instance embeddings. We experiment with sparse Hopfield pooling layers using our proposed α-entmax and α-normmax transformations (§3), as well as structured Hopfield pooling layers using SparseMAP with k-subsets (§4), varying α and k in each case. Note that 1-entmax recovers Ramsauer et al. (2021) and 2-entmax recovers Hu et al. (2023). We run these models for K-MIL problems in the MNIST dataset (choosing '9' as target) and in three MIL benchmarks: Elephant, Fox, and Tiger (Ilse et al., 2018). Further details can be found in App. F.1 and F.2.

Tab. 2 shows the results. We observe that for MNIST, $K = 1$, 1-entmax surprisingly outperforms the remaining methods. Normmax shows consistent results across datasets achieving near optimal performance, arguably due to its ability to adapt to near-uniform metastable states of varying size. We also observe that, for $K > 1$, the k-subsets approach achieves top performance when $k = K$, as expected. We also see that, in the MIL benchmarks, SparseMAP pooling surpasses sparse pooling variants for 2 out of 3 datasets.

5.4. Structured Rationalizers

We experiment with rationalizer models in sentiment prediction tasks, where the inputs are sentences or documents in natural language and the rationales are text highlights (see Fig. 4). These models, sometimes referred as select-predict or explain-predict models (Jacovi & Goldberg, 2021; Zhang et al., 2021), consist of a rationale generator and a predictor. The generator processes the input text and extracts the rationale as a subset of words to be highlighted, and the predictor classifies the input based solely on the extracted rationale, which generally involves concealing non-rationale words through the application of a binary mask. Rationalizers

Table 1: Distribution of metastable state (in %) in MNIST. The training set is memorized and the test set is used as queries.

Metastable State Size	$\beta = 0.1$								$\beta = 1$							
	α-entmax			α-normmax		k-subsets			α-entmax			α-normmax		k-subsets		
	1	1.5	2	2	5	2	4	8	1	1.5	2	2	5	2	4	8
1	3.5	69.2	88.1	81.4	51.4	0.0	0.0	0.0	97.8	99.9	100.0	100.0	99.8	0.0	0.0	0.0
2	2.1	8.6	5.2	6.7	31.4	87.3	0.0	0.0	0.9	0.1	0.0	0.0	0.2	99.9	0.0	0.0
3	1.6	3.9	2.6	1.9	7.0	6.1	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.1	0.0	0.0
4	1.2	2.3	1.6	1.0	2.1	2.5	80.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	99.3	0.0
5	1.2	1.6	1.1	0.9	1.5	2.0	11.9	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.7	0.0
6	0.9	0.9	0.8	0.5	1.5	1.1	4.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
7	1.1	0.6	0.4	0.4	1.3	0.6	2.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.8	0.6	0.1	0.8	1.0	0.2	1.0	60.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	95.0
9	1.0	0.3	0.0	0.5	0.8	0.1	0.4	26.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	4.7
10	1.1	0.1	0.0	0.5	0.6	0.0	0.1	9.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2
10^+	85.5	11.9	0.1	5.4	1.4	0.1	0.0	4.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1

Table 2: Results for MIL. We show accuracies for MNIST and ROC AUC for MIL benchmarks, averaged across 5 runs.

	MNIST				MIL benchmarks		
Methods	K=1	K=2	K=3	K=5	Fox	Tiger	Elephant
1-entmax (softmax)	$\mathbf{98.4 \pm 0.2}$	94.6 ± 0.5	91.1 ± 0.5	89.0 ± 0.3	66.4 ± 2.0	87.1 ± 1.6	92.6 ± 0.6
1.5-entmax	97.6 ± 0.8	96.0 ± 0.9	90.4 ± 1.1	92.4 ± 1.4	66.3 ± 2.0	87.3 ± 1.5	92.4 ± 1.0
2.0-entmax (sparsemax)	97.9 ± 0.2	96.7 ± 0.5	92.9 ± 0.9	91.6 ± 1.0	66.1 ± 0.6	$\mathbf{87.7 \pm 1.4}$	91.8 ± 0.6
2.0-normmax	97.9 ± 0.3	96.6 ± 0.6	93.9 ± 0.7	92.4 ± 0.7	66.1 ± 2.5	86.4 ± 0.8	92.4 ± 0.7
5.0-normmax	98.2 ± 0.5	97.2 ± 0.3	95.8 ± 0.4	93.2 ± 0.5	66.4 ± 2.3	85.5 ± 0.6	93.0 ± 0.7
SparseMAP, $k = 2$	97.9 ± 0.3	$\mathbf{97.7 \pm 0.3}$	95.1 ± 0.5	92.6 ± 1.1	66.8 ± 2.7	85.3 ± 0.5	$\mathbf{93.2 \pm 0.7}$
SparseMAP, $k = 3$	98.0 ± 0.6	96.1 ± 1.0	$\mathbf{96.5 \pm 0.5}$	92.2 ± 1.2	$\mathbf{67.4 \pm 2.0}$	86.1 ± 0.8	92.6 ± 1.7
SparseMAP, $k = 5$	98.2 ± 0.4	96.2 ± 1.4	95.1 ± 1.1	$\mathbf{95.1 \pm 1.5}$	67.0 ± 2.0	86.3 ± 0.8	91.2 ± 1.0

are usually trained end-to-end, and the discreteness of the latent rationales is either handled with stochastic methods via score function estimators or the reparametrization trick (Lei et al., 2016; Bastings et al., 2019), or with deterministic methods via structured continuous relaxations (Guerreiro & Martins, 2021). In either case, the model imposes sparsity and contiguity penalties to ensure rationales are short and tend to extract adjacent words. Our model architecture is adapted from SPECTRA (Guerreiro & Martins, 2021), but the combination of the generator and predictor departs from prior approaches (Lei et al., 2016; Bastings et al., 2019; Guerreiro & Martins, 2021) in which the predictor does not "mask" the input tokens; instead, it takes as input the pooled vector that results from the Hopfield pooling layer (either a sequential or non-sequential SparseMAP k-subsets layer). By integrating this Hopfield pooling layer into the predictor, we transform the sequence of word embeddings into a single vector from which the prediction is made. The rationale is formed by the pattern associations (word tokens) extracted by the Hopfield layer.

Tab. 3 shows the results in the downstream task (classification for SST, AgNews, IMDB; regression for BeerAdvocate) as well as the F_1 overlap with human rationales for the BeerAdvocate dataset (McAuley et al., 2012). Compared to strong baselines (Bastings et al., 2019; Guerreiro & Martins, 2021), our proposed methods achieve equal or slightly superior performance for all datasets. Moreover, our sequential k-subsets model outperforms the baselines in terms of overlap with human rationales. This is explained by the fact that human rationales tend to contain adjacent words, which is encouraged by our sequential model. Additional details and results with other baselines are shown in App. F.3.

6. Related Work

Recent research on modern Hopfield networks and dense associative memories includes (Krotov & Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2021; Millidge et al., 2022; Hoover et al., 2023, *inter alia*). The closest to our work is Hu et al. (2023), who proposed sparse Hopfield networks (equivalent to our 2-entmax, i.e., sparsemax) and derived retrieval error bounds tighter than the dense analog. Concurrently to our work, Wu et al. (2024) further proposed a "generalized sparse Hopfield model", which corresponds to α-entmax with learnable α, and applied it successfully to time series prediction problems. However, neither Hu et al. (2023) or Wu et al. (2024) seem to have realized the possibility of **exact** retrieval enabled by sparse transformations. Our paper fills this gap by providing a unified framework for sparse Hopfield networks with stronger theoretical guarantees for retrieval and coverage (cf. Prop. 2–3). Our

Table 3: Text rationalization results. We report mean and min/max F_1 scores across five random seeds on test sets for all datasets but Beer, where we report MSE. HardKuma and SPECTRA results are taken from (Guerreiro & Martins, 2021). We also report human rationale overlap (HRO) as F_1 score. We bold the best-performing system(s).

	SST↑	AgNews↑	IMDB↑	Beer (MSE) ↓	Beer (HRO) ↑
HardKuma (Bastings et al., 2019)	.80 (.80/.81)	.90 (.87/.88)	.87 (.90/.91)	.019 (.016/.020)	.37 (.00/.90)
SPECTRA (Guerreiro & Martins, 2021)	.80 (.79/.81)	.92 (.92/.93)	**.90** (.89/.90)	**.017** (.016/.019)	.61 (.56/.68)
SparseMAP k-subsets (ours)	**.81** (.81/.82)	**.93** (.92/.93)	**.90** (.90/.90)	**.017** (.017/.018)	.42 (.29/.62)
SparseMAP seq. k-subsets (ours)	**.81** (.80/.83)	**.93** (.93/.93)	**.90** (.90/.90)	.020 (.018/.021)	**.63** (.49/.70)

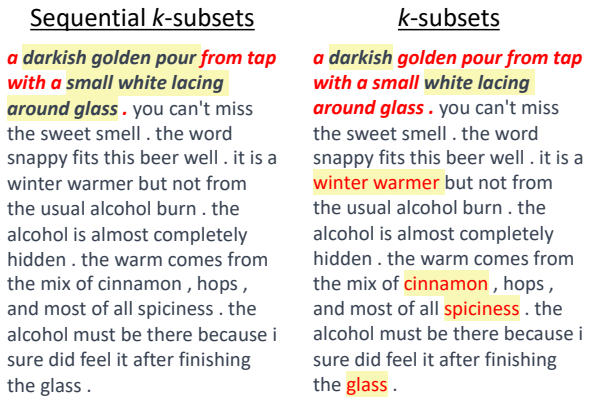

Figure 4: Example of human rationale overlap for the aspect "appearance". The yellow highlight indicates the model's rationale, while ***italicized and bold font*** represents the human rationale. Red font identifies mismatches with human annotations. SparseMAP with sequential k-subsets prefers more contiguous rationales, which better match humans. Additional examples are shown in App. F.3.

framework generalizes their constructions and widens the scope to new families, such as α-normmax, for which we provide an effective bisection algorithm (Alg. 1 in App. B).

The link with Fenchel-Young losses (Blondel et al., 2020) is a key dimension of our work. Many results derived therein, such as the margin conditions, were found to have direct application to sparse Hopfield networks. In our paper, we extend some of their results, such as the structured margin of SparseMAP (Prop. 4), key to establish exact retrieval of pattern associations (Prop. 5). Our k-subsets example relates to the top-k retrieval model of Davydov et al. (2023): their model differs from ours as it uses an entropic regularizer not amenable to sparsity, making exact retrieval impossible.

Our sparse and structured Hopfield layers with sparsemax and SparseMAP involve a quadratic regularizer, which relates to the differentiable layers of Amos & Kolter (2017). The use of SparseMAP and its active set algorithm (Niculae et al., 2018) allows to exploit the structure of the problem to ensure efficient Hopfield updates and implicit derivatives.

7. Conclusions

We presented a unified framework for sparse and structured Hopfield networks. Our framework hinges on a broad family of energy functions written as a difference of a quadratic regularizer and a Fenchel-Young loss, parametrized by a generalized negentropy function. A core result of our paper is the link between the margin property of certain Fenchel-Young losses and sparse Hopfield networks with provable conditions for exact retrieval. We further extended this framework to incorporate structure via the SparseMAP transformation, which is able to retrieve pattern associations instead of a single pattern. Empirical evaluation confirmed the usefulness of our approach in multiple instance learning and text rationalization problems.

Impact Statement

Hopfield networks are increasingly relevant for practical applications and not only as theoretical models. While our work is mostly a theoretical advancement, promising experimental results signal potentially wider impact. Sparse HFY networks are applicable in the same scenarios where modern Hopfield networks would be, and we do not foresee any specific societal consequences of sparse transformations or exact retrieval in such cases. In the structured case, the practicioner has more hands-on control for encoding inductive biases through the design of a factor graph and the choice of its parameters (e.g., k). These choices may in turn reflect human biases and have societal implications: for example, contiguous rationales might be well suited for English but not for other languages. We encourage care when designing such methods for practical 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

Amit, D. J., Gutfreund, H., and Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

Amos, B. and Kolter, J. Z. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pp. 136–145. PMLR, 2017.

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

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

Brauchart, J. S., Reznikov, A. B., Saff, E. B., Sloan, I. H., Wang, Y. G., and Womersley, R. S. Random point sets on the sphere—hole radii, covering, and separation. *Experimental Mathematics*, 27(1):62–81, 2018.

Chabauty, C. Resultats sur l'empilement de calottes egales sur une perisphere de $\mathbb{R}^n$ et correction a un travail anterieur. *Comptes Rendus Hebdomadaires des Seances de l'Academie des Sciences*, 236(15):1462–1464, 1953.

Conway, J. H. and Sloane, N. J. A. *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media, 2013.

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

Davydov, A., Jaffe, S., Singh, A., and Bullo, F. Retrieving k-nearest memories with modern Hopfield networks. In *Associative Memory & Hopfield Networks in 2023*, 2023.

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

Guerreiro, N. M. and Martins, A. F. T. Spectra: Sparse structured text rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6534–6550, 2021.

Hoover, B., Liang, Y., Pham, B., Panda, R., Strobelt, H., Chau, D. H., Zaki, M. J., and Krotov, D. Energy transformer. In *Advances in Neural Information Processing Systems*, 2023.

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

Hu, J. Y.-C., Yang, D., Wu, D., Xu, C., Chen, B.-Y., and Liu, H. On sparse modern hopfield model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2309.12673.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2127–2136. PMLR, 2018. URL https://arxiv.org/abs/1802.04712.

Jacovi, A. and Goldberg, Y. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310, 2021.

Jenssen, M., Joos, F., and Perkins, W. On kissing numbers and spherical codes in high dimensions. *Advances in Mathematics*, 335:307–321, 2018.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, 2016.

Kschischang, F. R., Frey, B. J., and Loeliger, H.-A. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

Lei, T., Barzilay, R., and Jaakkola, T. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016.

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

Li, Z. and Arora, S. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJg8TeSFDH.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

May, R. M. Patterns of species abundance and distribution. *Ecology and Evolution of Communities*, pp. 81–120, 1975.

McAuley, J. J., Leskovec, J., and Jurafsky, D. Learning attitudes and attributes from multi-aspect reviews. In *ICDM*, 2012.

McEliece, R., Posner, E., Rodemich, E., and Venkatesh, S. The capacity of the hopfield associative memory. *IEEE transactions on Information Theory*, 33(4):461–482, 1987.

Millidge, B., Salvatori, T., Song, Y., Lukasiewicz, T., and Bogacz, R. Universal hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, pp. 15561–15583. PMLR, 2022.

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

Niculae, V. and Blondel, M. A regularized framework for sparse and structured neural attention. *Advances in Neural Information Processing Systems*, 30, 2017.

Niculae, V., Martins, A. F., Blondel, M., and Cardie, C. Sparsemap: Differentiable sparse structured inference. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

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

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

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. Hopfield networks is all you need. In *Proceedings of ICLR*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

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

Shannon, C. E. Probability of error for optimal codes in a gaussian channel. *Bell System Technical Journal*, 38(3): 611–656, 1959.

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

Taskar, B., Guestrin, C., and Koller, D. Max-margin markov networks. In *Advances in Neural Information Processing Systems*, 2003.

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

Tse, D., Langston, R., Kakeyama, M., Bethus, I., Spooner, P., Wood, E., Witter, M., and Morris, R. Schemas and memory consolidation. *Science*, 316(5821):76–82, 2007. doi: 10.1126/science.1135935.

Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., and Singer, Y. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(9), 2005.

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

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

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

Wu, D., Hu, J. Y.-C., Li, W., Chen, B.-Y., and Liu, H. STanhop: Sparse tandem hopfield model for memory-enhanced time series prediction. In *Proceedings of ICLR*, 2024. URL https://openreview.net/forum?id=6iwg437CZs.

Wyner, A. D. Capabilities of bounded discrepancy decoding. *Bell System Technical Journal*, 44(6):1061–1122, 1965.

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

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

Zhang, Z., Rudra, K., and Anand, A. Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021.

A. Generalized Negentropies

We recall here the definition of generalized negenetropies from Blondel et al. (2020, §4.1):

Definition 4. *A function $\Omega : \triangle_N \to \mathbb{R}$ is a generalized negentropy iff it satisfies the following conditions:*

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

This definition implies that $\Omega \leq 0$ and that Ω is minimized when $\boldsymbol{y} = \boldsymbol{1}/N$ is the uniform distribution (Blondel et al., 2020, Prop. 4). This justifies the name "generalized negentropies."

B. Bisection Algorithm for the Normmax Transformation

We derive here expressions for the normmax transformation along with a bisection algorithm to compute this transformation for general α.

Letting $\Omega(\boldsymbol{y}) = -1 + \|\boldsymbol{y}\|_\alpha + I_{\triangle_N}(\boldsymbol{y})$ be the norm entropy, we have

$$(\nabla\Omega^*)(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{y}\in\triangle_N} \boldsymbol{\theta}^\top \boldsymbol{y} - \|\boldsymbol{y}\|_q. \tag{14}$$

The Lagrangian function is $L(\boldsymbol{y}, \boldsymbol{\lambda}, \mu) = -\boldsymbol{\theta}^\top \boldsymbol{y} + \|\boldsymbol{y}\|_\alpha - \boldsymbol{\lambda}^\top \boldsymbol{y} + \mu(\boldsymbol{1}^\top \boldsymbol{y} - 1)$. Equating the gradient to zero and using the fact that $\nabla\|\boldsymbol{y}\|_\alpha = \left(\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_\alpha}\right)^{\alpha-1}$, we get:

$$\boldsymbol{0} = \nabla_{\boldsymbol{y}} L(\boldsymbol{y}, \boldsymbol{\lambda}, \mu) = -\boldsymbol{\theta} + \left(\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_\alpha}\right)^{\alpha-1} - \boldsymbol{\lambda} + \mu\boldsymbol{1}. \tag{15}$$

The complementarity slackness condition implies that, if $y_i > 0$, we must have $\lambda_i = 0$, therefore, we have for such $i \in \operatorname{supp}(\boldsymbol{y})$:

$$-\theta_i + \left(\frac{y_i}{\|\boldsymbol{y}\|_\alpha}\right)^{\alpha-1} + \mu = 0 \quad \Rightarrow \quad y_i = (\theta_i - \mu)^{\frac{1}{\alpha-1}} \|\boldsymbol{y}\|_\alpha. \tag{16}$$

Since we must have $\sum_{i\in\operatorname{supp}(\boldsymbol{y})} y_i = 1$, we obtain:

$$1 = \sum_{i\in\operatorname{supp}(\boldsymbol{y})} (\theta_i - \mu)^{\frac{1}{\alpha-1}} \|\boldsymbol{y}\|_\alpha \quad \Longrightarrow \quad \|\boldsymbol{y}\|_\alpha = \frac{1}{\sum_{i\in\operatorname{supp}(\boldsymbol{y})} (\theta_i - \mu)^{\frac{1}{\alpha-1}}}. \tag{17}$$

Combining the two last equations and noting that, from (15), we have $\theta_i < \mu_i$ for $i \notin \operatorname{supp}(\boldsymbol{y})$, we get, for $i \in [N]$:

$$y_i = \frac{(\theta_i - \mu)_+^{\frac{1}{\alpha-1}}}{\sum_{j\in\operatorname{supp}(\boldsymbol{y})} (\theta_j - \mu)_+^{\frac{1}{\alpha-1}}}. \tag{18}$$

Moreover, since $\sum_{i\in\operatorname{supp}(\boldsymbol{y})} y_i^\alpha = \|\boldsymbol{y}\|_\alpha^\alpha$, we obtain from (16):

$$\|\boldsymbol{y}\|_\alpha^\alpha = \sum_{i\in\operatorname{supp}(\boldsymbol{y})} (\theta_i - \mu)^{\frac{\alpha}{\alpha-1}} \|\boldsymbol{y}\|_\alpha^\alpha \quad \Rightarrow \quad \sum_{i\in\operatorname{supp}(\boldsymbol{y})} (\theta_i - \mu)^{\frac{\alpha}{\alpha-1}} = 1. \tag{19}$$

In order to compute the solution (18) we need to find μ satisfying (19). This can be done with a simple bisection algorithm if we find a lower and upper bound on μ.

We have, from (18), that $\mu = \theta_i - (y_i/\|\boldsymbol{y}\|_\alpha)^{\alpha-1}$ for any $i \in \operatorname{supp}(\boldsymbol{y})$. Letting $\theta_{\max} = \max_i \theta_i$ and $y_{\max} = \max_i y_i$, we have in particular that $\mu = \theta_{\max} - (y_{\max}/\|\boldsymbol{y}\|_\alpha)^{\alpha-1}$. We also have that $y_{\max} = \|\boldsymbol{y}\|_\infty \leq \|\boldsymbol{y}\|_\alpha$, which implies

Algorithm 1 Compute α-normmax by bisection.

Input: Scores $\boldsymbol{\theta} = [\theta_1, ..., \theta_N]^\top \in \mathbb{R}^N$, parameter $\alpha > 1$.
Output: Probability vector $\boldsymbol{y} = [y_1, ..., y_N]^\top \in \triangle_N$.
Define $\theta_{\max} \leftarrow \max_i \theta_i$
Compute $\mu_{\min} \leftarrow \theta_{\max} - 1$
Compute $\mu_{\max} \leftarrow \theta_{\max} - N^{1-\alpha}$
for $t \in 1, \ldots, T$ **do**
  $\mu \leftarrow (\mu_{\min} + \mu_{\max})/2$
  $Z \leftarrow \sum_j (\theta_j - \mu)_+^{\frac{\alpha}{\alpha-1}}$
  if $Z < 1$ **then** $\mu_{\max} \leftarrow \mu$ **else** $\mu_{\min} \leftarrow \mu$
end for
Return $\boldsymbol{y} = [y_1, ..., y_N]^\top$ with $y_i = \frac{(\theta_i - \mu)_+^{\frac{1}{\alpha-1}}}{\sum_j (\theta_j - \mu)_+^{\frac{1}{\alpha-1}}}$.

$y_{\max}/\|\boldsymbol{y}\|_\alpha \leq 1$. Since $1/N \leq y_{\max} \leq 1$ and $\|\boldsymbol{y}\|_\alpha \leq 1$ for any $\boldsymbol{y} \in \triangle_N$, we also obtain $y_{\max}/\|\boldsymbol{y}\|_\alpha \geq (1/N)/1 = N^{-1}$. Therefore we have

$$\underbrace{\theta_{\max} - 1}_{\mu_{\min}} \leq \mu \leq \underbrace{\theta_{\max} - N^{1-\alpha}}_{\mu_{\max}}. \tag{20}$$

The resulting algorithm is shown as Alg. 1.

C. Structured Prediction with Factor Graphs

We define our notation and setup for structured prediction with factor graph representations, based on Niculae et al. (2018).

We assume the interactions among patterns (for example due to temporal dependencies, hierarchical or link structure, etc.) can be expressed as a bipartite **factor graph** (V, F), where V is a set of variable nodes and F are factor nodes (Kschischang et al., 2001). Each factor $f \in F$ is linked to a subset of variable nodes $V_f \subseteq V$. We assume each variable $v \in V$ can take one of N_v possible values, and we denote by $\boldsymbol{y}_v \in \{0,1\}^{N_v}$ a one-hot vector indicating a value for this variable. Likewise, each factor $f \in F$ has N_f possible configurations, with $N_f = \prod_{v \in V_f} N_v$, and we associate to it a one-hot vector $\boldsymbol{y}_f \in \{0,1\}^{N_f}$ indicating a configuration for that factor. The global configuration of the factor graph is expressed through the bit vectors $\boldsymbol{y}_V = [\boldsymbol{y}_v : v \in V] \in \{0,1\}^{N_V}$ and $\boldsymbol{y}_F = [\boldsymbol{y}_f : f \in F] \in \{0,1\}^{N_F}$, with $N_V = \sum_{v \in V} N_v$ and $N_F = \sum_{f \in F} N_f$. A particular structure is expressed through the bit vector $\boldsymbol{y} = [\boldsymbol{y}_V; \boldsymbol{y}_F] \in \{0,1\}^{N_V + N_F}$. Finally, we define the set of **valid structures** $\mathcal{Y} \subseteq \{0,1\}^{N_V + N_F}$ – this set contains all the bit vectors which correspond to valid structures, which must satisfy consistency between variable and factor assignments, as well as any additional hard constraints.

We associate **unary scores** $\boldsymbol{\theta}_V = [\boldsymbol{\theta}_v : v \in V] \in \mathbb{R}^{N_V}$ to configurations of variable nodes and **higher-order scores** $\boldsymbol{\theta}_F = [\boldsymbol{\theta}_f : f \in F] \in \mathbb{R}^{N_F}$ to configurations of factor nodes. We denote $\boldsymbol{\theta} = [\boldsymbol{\theta}_V; \boldsymbol{\theta}_F] \in \mathbb{R}^{N_V + N_F}$. The problem of finding the highest-scoring structure, called the **maximum a posteriori** (MAP) inference problem, is

$$\boldsymbol{y}^* = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{\theta}^\top \boldsymbol{y} = \operatorname*{argmax}_{\boldsymbol{y} \in \operatorname{conv}(\mathcal{Y})} \boldsymbol{\theta}_V^\top \boldsymbol{y}_V + \boldsymbol{\theta}_F^\top \boldsymbol{y}_F. \tag{21}$$

As above, we consider regularized variants of (21) via a convex regularizer $\Omega : \operatorname{conv}(\mathcal{Y}) \to \mathbb{R}$. **SparseMAP** corresponds to $\Omega(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}_V\|^2$ (note that only the unary variables are regularized), which leads to the quadratic optimization problem (12). Niculae et al. (2018) developed an effective and computationally efficient active set algorithm for solving this problem, which requires only a MAP oracle to solve instances of the problem (21).

D. Proofs of Main Text

D.1. 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

$$L_\Omega(\boldsymbol{\theta}; \mathbf{1}/N) \leq \max_i \theta_i - \mathbf{1}^\top \boldsymbol{\theta}/N. \tag{22}$$

From the definition of Ω^* and the fact that $\Omega(\boldsymbol{y}) \geq \Omega(\mathbf{1}/N)$ for any $\boldsymbol{y} \in \triangle_N$, we have that, for any $\boldsymbol{\theta}$, $\Omega^*(\boldsymbol{\theta}) = \max_{\boldsymbol{y}\in\triangle_N} \boldsymbol{\theta}^\top \boldsymbol{y} - \Omega(\boldsymbol{y}) \leq \max_{\boldsymbol{y}\in\triangle_N} \boldsymbol{\theta}^\top \boldsymbol{y} - \Omega(\mathbf{1}/N) = \max_i \theta_i - \Omega(\mathbf{1}/N)$, which leads to (22).

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 $\boldsymbol{q}$. 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 $\boldsymbol{y} \in \triangle_N$, we have $0 \leq L_\Omega(\boldsymbol{\theta}, \boldsymbol{y}) = L_\Omega(\boldsymbol{\theta}, \mathbf{1}/N) - \Omega(\mathbf{1}/N) + \Omega(\boldsymbol{y}) - (\boldsymbol{y} - \mathbf{1}/N)^\top \boldsymbol{\theta} \leq L_\Omega(\boldsymbol{\theta}, \mathbf{1}/N) - \Omega(\mathbf{1}/N) - (\boldsymbol{y} - \mathbf{1}/N)^\top \boldsymbol{\theta}$, due to the assumptions 1–3 which ensure Ω is non-positive. That is, $L_\Omega(\boldsymbol{\theta}, \mathbf{1}/N) \geq \Omega(\mathbf{1}/N) + (\boldsymbol{y} - \mathbf{1}/N)^\top \boldsymbol{\theta}$. Therefore, with $\boldsymbol{q} = \boldsymbol{X}^\top \boldsymbol{y}$, we get

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

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 $\|\boldsymbol{q}\| = \|\boldsymbol{X}^\top \boldsymbol{y}\| \leq \sum_i y_i \|x_i\| \leq M$ 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 tth 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 $\nabla L_\Omega(\boldsymbol{\theta}, \boldsymbol{y}) = \hat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$ and the chain rule leads to

$$\begin{aligned}
\nabla E_{\text{concave}}(\boldsymbol{q}) &= -\beta^{-1}\nabla_{\boldsymbol{q}} L_\Omega(\beta \boldsymbol{X}\boldsymbol{q}; \mathbf{1}/N) = \boldsymbol{X}^\top(\mathbf{1}/N - \hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q})) \\
&= \boldsymbol{\mu}_{\boldsymbol{X}} - \boldsymbol{X}^\top \hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q}) \\
\nabla E_{\text{convex}}(\boldsymbol{q}) &= \boldsymbol{q} - \boldsymbol{\mu}_{\boldsymbol{X}},
\end{aligned} \tag{23}$$

giving the update equation (9).

D.2. 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 (23), this is equivalent to $\boldsymbol{q} = \boldsymbol{X}^\top \hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q})$. 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 $\hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{x}_i) = \boldsymbol{e}_i$. We now use the margin property of sparse transformations (7), according to which the latter is equivalent to $\beta \boldsymbol{x}_i^\top \boldsymbol{x}_i - \max_{j\neq i} \beta \boldsymbol{x}_i^\top \boldsymbol{x}_j \geq m$. Noting that the left hand side equals $\beta\Delta_i$ leads to the desired result.

If the initial query satisfies ${\boldsymbol{q}^{(0)}}^\top(\boldsymbol{x}_i - \boldsymbol{x_j}) \geq \frac{m}{\beta}$ for all $j \neq i$, we have again from the margin property that $\hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q}^{(0)}) = \boldsymbol{e}_i$, 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

$$\begin{aligned}(\boldsymbol{q}^{(0)})^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) &= (\boldsymbol{x}_i + \epsilon \boldsymbol{r})^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &\geq \Delta_i + \epsilon \boldsymbol{r}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &\geq \Delta_i - \epsilon \underbrace{\|\boldsymbol{r}\|}_{=1} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|, \end{aligned} \tag{24}$$

where we invoked the Cauchy-Schwarz inequality in the last step. Since the patterns are normalized (with norm M),[4] we have from the triangle inequality that $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq \|\boldsymbol{x}_i\| + \|\boldsymbol{x}_j\| = 2M$; using the assumption that $\Delta_i \geq \frac{m}{\beta} + 2M\epsilon$, we obtain ${\boldsymbol{q}^{(0)}}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq \frac{m}{\beta}$, which from the previous points ensures convergence to $\boldsymbol{x}_i$ in one iteration.

D.3. Proof of Proposition 3

For the first statement, we follow a similar argument as the one made by Ramsauer et al. (2021) in the proof of their Theorem A.3—however their proof has a mistake, which we correct here.[5] Given a separation angle $\alpha_{\min}$, we lower bound the number of patterns N we can can place in the sphere separated by at least this angle. Estimating this quantity is an important open problem in combinatorics, related to determining the size of spherical codes (of which kissing numbers are a particular case; Conway & Sloane 2013). We invoke a lower bound due to Chabauty (1953), Shannon (1959), and Wyner (1965) (see also Jenssen et al. (2018) for a tighter bound), who show that $N \geq (1 + o(1))\sqrt{2\pi D}\frac{\cos\alpha_{\min}}{(\sin\alpha_{\min})^{D-1}}$. For $\alpha_{\min} = \frac{\pi}{3}$, which corresponds to the kissing number problem, we obtain the bound

$$N \geq (1 + o(1))\sqrt{\frac{3\pi D}{8}}\left(\frac{2}{\sqrt{3}}\right)^D = \mathcal{O}\left(\left(\frac{2}{\sqrt{3}}\right)^D\right).$$

In this scenario, we have $\Delta_i = M^2(1 - \cos\alpha_{\min})$ by the definition of Δ_i. From Proposition 2, we have exact retrieval under ϵ-perturbations if $\Delta_i \geq m\beta^{-1} + 2M\epsilon$. Combining the two expressions, we obtain $\epsilon \leq \frac{M}{2}(1 - \cos\alpha_{\min}) - \frac{m}{2\beta M}$. Setting $\alpha_{\min} = \frac{\pi}{3}$, we obtain $\epsilon \leq \frac{M}{2}\left(1 - \frac{1}{2}\right) - \frac{m}{2\beta M} = \frac{M}{4} - \frac{m}{2\beta M}$. For the right hand side to be positive, we must have $M^2 > 2m/\beta$.

Assume now patterns are placed uniformly at random in the sphere. From Brauchart et al. (2018) we have, for any $\delta > 0$:

$$P(N^{\frac{2}{D-1}}\alpha_{\min} \geq \delta) \geq 1 - \frac{\kappa_{D-1}}{2}\delta^{D-1}, \quad \text{with} \quad \kappa_D := \frac{1}{D\sqrt{\pi}}\frac{\Gamma((D+1)/2)}{\Gamma(D/2)}. \tag{25}$$

Given our failure probability p, we need to have

$$P(M^2(1 - \cos\alpha_{\min}) \geq m\beta^{-1} + 2M\epsilon) \geq 1 - p, \tag{26}$$

which is equivalent to

$$P\left\{N^{\frac{2}{D-1}}\alpha_{\min} \geq \underbrace{N^{\frac{2}{D-1}}\arccos\left(1 - \frac{m}{\beta M^2} - \frac{2\epsilon}{M}\right)}_{:=\delta}\right\} \geq 1 - p. \tag{27}$$

Therefore, we set

$$p = \frac{\kappa_{D-1}}{2}\delta^{D-1} = \frac{\kappa_{D-1}}{2}N^2\left[\arccos\left(1 - \frac{m}{\beta M^2} - \frac{2\epsilon}{M}\right)\right]^{D-1}. \tag{28}$$

[4] In fact, the result still holds if patterns are not normalized but have their norm upper bounded by M, i.e., if they lie within a ball of radius M and not necessarily on the sphere.

[5] Concretely, Ramsauer et al. (2021) claim that given a separation angle $\alpha_{\min}$, we can place $N = (2\pi/\alpha_{\min})^{D-1}$ patterns equidistant on the sphere, but this is not correct.

Choosing $N = \sqrt{\frac{2p}{\kappa_{D-1}}}\zeta^{\frac{D-1}{2}}$ patterns for some $\zeta > 1$, we obtain

$$1 = \left[\zeta \arccos\left(1 - \frac{m}{\beta M^2} - \frac{2\epsilon}{M}\right)\right]^{D-1}. \tag{29}$$

Therefore the failure rate p is attainable provided the perturbation error is

$$\epsilon \leq \frac{M}{2}\left(1 - \cos\frac{1}{\zeta}\right) - \frac{m}{2\beta M}. \tag{30}$$

For the right hand side to be positive, we must have $\cos\frac{1}{\zeta} < 1 - \frac{m}{\beta M^2}$, i.e., $\zeta < \frac{1}{\arccos\left(1-\frac{m}{\beta M^2}\right)}$.

D.4. Proof of Proposition 4

The first statement in the proposition is stated and proved by Blondel et al. (2020) as a corollary of their Proposition 8. We prove here a more general version, which includes the second statement as a novel result.

Using Blondel et al. (2020, Proposition 8), we have that the structured margin of L_Ω is given by the following expression,

$$m = \sup_{\substack{\boldsymbol{y}\in\mathcal{Y} \\ \boldsymbol{\mu}\in\operatorname{conv}(\mathcal{Y})}} \frac{\Omega(\boldsymbol{y}) - \Omega(\boldsymbol{\mu})}{r^2 - \boldsymbol{\mu}^\top \boldsymbol{y}},$$

if the supremum exists. For SparseMAP, using $\Omega(\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\mu}_V\|^2 = \frac{1}{2}\|\boldsymbol{\mu}\|^2 - \frac{1}{2}\|\boldsymbol{\mu}_F\|^2$ for any $\boldsymbol{\mu} \in \operatorname{conv}(\mathcal{Y})$, and using the fact that $\|\boldsymbol{y}\| = r$ for any $\boldsymbol{y} \in \mathcal{Y}$, we obtain:

$$\begin{aligned}
m &= \sup_{\substack{\boldsymbol{y}\in\mathcal{Y} \\ \boldsymbol{\mu}\in\operatorname{conv}(\mathcal{Y})}} \frac{\frac{1}{2}r^2 - \frac{1}{2}\|\boldsymbol{\mu}\|^2 + \overbrace{\frac{1}{2}\|\boldsymbol{\mu}_F\|^2 - \frac{1}{2}r_F^2}^{\leq 0}}{\boldsymbol{y}^\top(\boldsymbol{y}-\boldsymbol{\mu})} \\
&\leq^{(\dagger)} \sup_{\substack{\boldsymbol{y}\in\mathcal{Y} \\ \boldsymbol{\mu}\in\operatorname{conv}(\mathcal{Y})}} \frac{\frac{1}{2}r^2 - \frac{1}{2}\|\boldsymbol{\mu}\|^2}{\boldsymbol{y}^\top(\boldsymbol{y}-\boldsymbol{\mu})} \\
&= 1 - \inf_{\substack{\boldsymbol{y}\in\mathcal{Y} \\ \boldsymbol{\mu}\in\operatorname{conv}(\mathcal{Y})}} \frac{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{\boldsymbol{y}^\top(\boldsymbol{y}-\boldsymbol{\mu})} \\
&\leq^{(\ddagger)} 1,
\end{aligned}$$

where the inequality $^{(\dagger)}$ follows from the convexity of $\frac{1}{2}\|\cdot\|^2$, which implies that $\frac{1}{2}\|\boldsymbol{\mu}_F\|^2 \leq \frac{1}{2}\|\boldsymbol{y}_F\|^2 = \frac{1}{2}r_F^2$; and the inequality $^{(\ddagger)}$ follows from the fact that both the numerator and denominator in the second term are non-negative, the latter due to the Cauchy-Schwartz inequality and the fact that $\|\boldsymbol{\mu}\| \leq r$. This proves the second part of Proposition 4.

To prove the first part, note first that, if there are no higher order interactions, then $r_F = 0$ and $\boldsymbol{\mu}_F$ is an "empty vector", which implies that $^{(\dagger)}$ is an equality. We prove now that, in this case, $^{(\ddagger)}$ is also an equality, which implies that $m = 1$. We do that by showing that, for any $\boldsymbol{y} \in \mathcal{Y}$, we have $\inf_{\boldsymbol{\mu}\in\operatorname{conv}(\mathcal{Y})} \frac{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{\boldsymbol{y}^\top(\boldsymbol{y}-\boldsymbol{\mu})} = 0$. Indeed, choosing $\boldsymbol{\mu} = t\boldsymbol{y}' + (1-t)\boldsymbol{y}$ for an arbitrary $\boldsymbol{y}' \in \mathcal{Y} \setminus \{\boldsymbol{y}\}$, and letting $t \to 0^+$, we obtain $\frac{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{\boldsymbol{y}^\top(\boldsymbol{y}-\boldsymbol{\mu})} = \frac{\frac{t}{2}\|\boldsymbol{y}-\boldsymbol{y}'\|^2}{\boldsymbol{y}^\top(\boldsymbol{y}-\boldsymbol{y}')} \to 0$.

D.5. Proof of Proposition 5

A point $\boldsymbol{q}$ is stationary iff it satisfies $\boldsymbol{q} = \boldsymbol{X}^\top \hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{q})$. Therefore, $\boldsymbol{X}^\top \boldsymbol{y}_i$ is guaranteed to be a stationary point if[6] $\hat{\boldsymbol{y}}_\Omega(\beta \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{y}_i) = \boldsymbol{y}_i$, which is equivalent to zero loss, i.e., to the existence of a margin $\underbrace{\beta \boldsymbol{y}_i^\top \boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j)}_{\geq \beta\Delta_i} \geq \frac{1}{2}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$

[6] But not necessarily "only if" – in general, we could have $\boldsymbol{X}^\top \boldsymbol{y}_i$ in the convex hull of the other pattern associations.

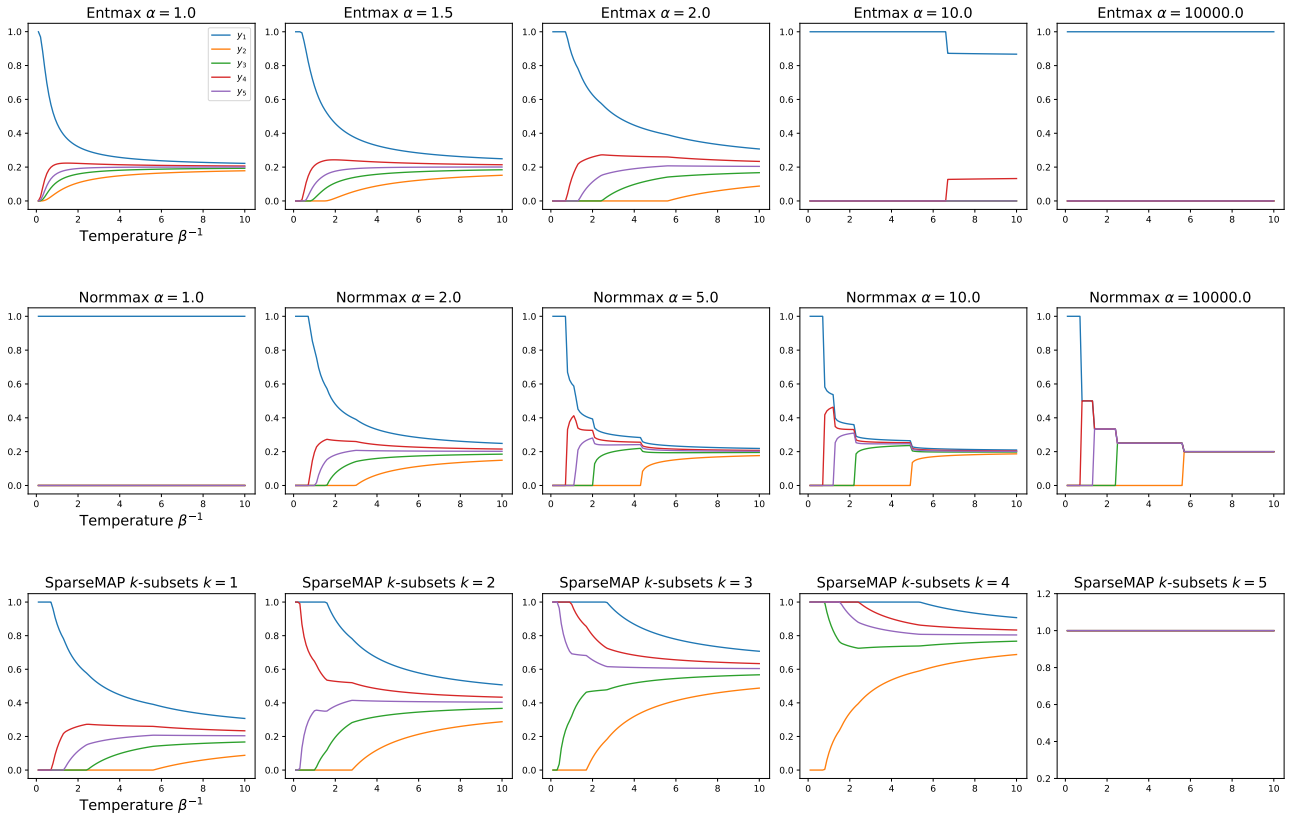

Figure 5: Sparse and structured transformations used in this paper and their regularization path. In each plot, we show $\hat{\boldsymbol{y}}_\Omega(\beta\boldsymbol{\theta}) = \hat{\boldsymbol{y}}_{\beta^{-1}\Omega}(\boldsymbol{\theta})$ as a function of the temperature β^{-1} where $\boldsymbol{\theta} = [1.0716, -1.1221, -0.3288, 0.3368, 0.0425]^\top$.

for all j. Since we are assuming $\Delta_i \geq \frac{D_i^2}{2\beta} \geq \frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{2\beta}$ for all j, we have $\beta\Delta_i \geq \frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{2}$ for all j, which implies the margin condition above.

If the initial query satisfies $\boldsymbol{q}^\top \boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) \geq \frac{D_i^2}{2\beta}$ for all $j \neq i$, we have again from the margin property that $\hat{\boldsymbol{y}}_\Omega(\beta\boldsymbol{X}\boldsymbol{q}) = \boldsymbol{y}_i$, which ensures convergence in one step to $\boldsymbol{X}^\top\boldsymbol{y}_i$.

If $\boldsymbol{q}$ is ϵ-close to $\boldsymbol{X}^\top\boldsymbol{y}_i$, then we have $\boldsymbol{q} = \boldsymbol{X}^\top\boldsymbol{y}_i + \epsilon\boldsymbol{r}$ for some vector $\boldsymbol{r}$ with $\|\boldsymbol{r}\| = 1$. Therefore, we have

$$\begin{aligned} \boldsymbol{q}^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) &= (\boldsymbol{X}^\top\boldsymbol{y}_i + \epsilon\boldsymbol{r})^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) \\ &\geq \Delta_i + \epsilon\boldsymbol{r}^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j). \end{aligned} \tag{31}$$

We now bound $-\boldsymbol{r}^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j)$ in two possible ways. Using the Cauchy-Schwarz inequality, we have $-\boldsymbol{r}^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) \leq \|\boldsymbol{X}\boldsymbol{r}\|\|\boldsymbol{y}_i - \boldsymbol{y}_j\| \leq \sigma_{\max}(\boldsymbol{X})D_i$, where $\sigma_{\max}(\boldsymbol{X})$ denotes the largest singular value of $\boldsymbol{X}$, i.e., the spectral norm of $\boldsymbol{X}$. On the other hand, denoting $R_i := \max_j \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_1$, we can also use Hölder's inequality to obtain $-\boldsymbol{r}^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) \leq \|\boldsymbol{X}\boldsymbol{r}\|_\infty\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_1 \leq MR_i$, where we used the fact that $\|\boldsymbol{X}\boldsymbol{r}\|_\infty = \max_k |\boldsymbol{x}_k^\top\boldsymbol{r}| \leq \|\boldsymbol{x}_k\|\|\boldsymbol{r}\| = M$. Combining the two inequalities, we obtain $\boldsymbol{q}^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) \geq \Delta_i - \epsilon\min\{\sigma_{\max}(\boldsymbol{X})D_i, MR_i\}$. Using the assumption that $\Delta_i \geq \frac{D_i^2}{2\beta} + \epsilon\min\{\sigma_{\max}(\boldsymbol{X})D_i, MR_i\}$, we obtain $\boldsymbol{q}^\top\boldsymbol{X}^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) \geq \frac{D_i^2}{2\beta}$, which from the previous points ensures convergence to $\boldsymbol{X}^\top\boldsymbol{y}_i$ in one iteration. The result follows by noting that, since $\mathcal{Y} \subseteq \{0,1\}^D$, we have $R_i = D_i^2$.

E. Additional Details and Experiments

E.1. Sparse and structured transformations

We show additional sparse and structured transformations and their regularization paths in Figure 5. The difference between the entmax and normmax regularizers is subtle but important: when $\alpha \to \infty$, the entmax regularizer vanishes

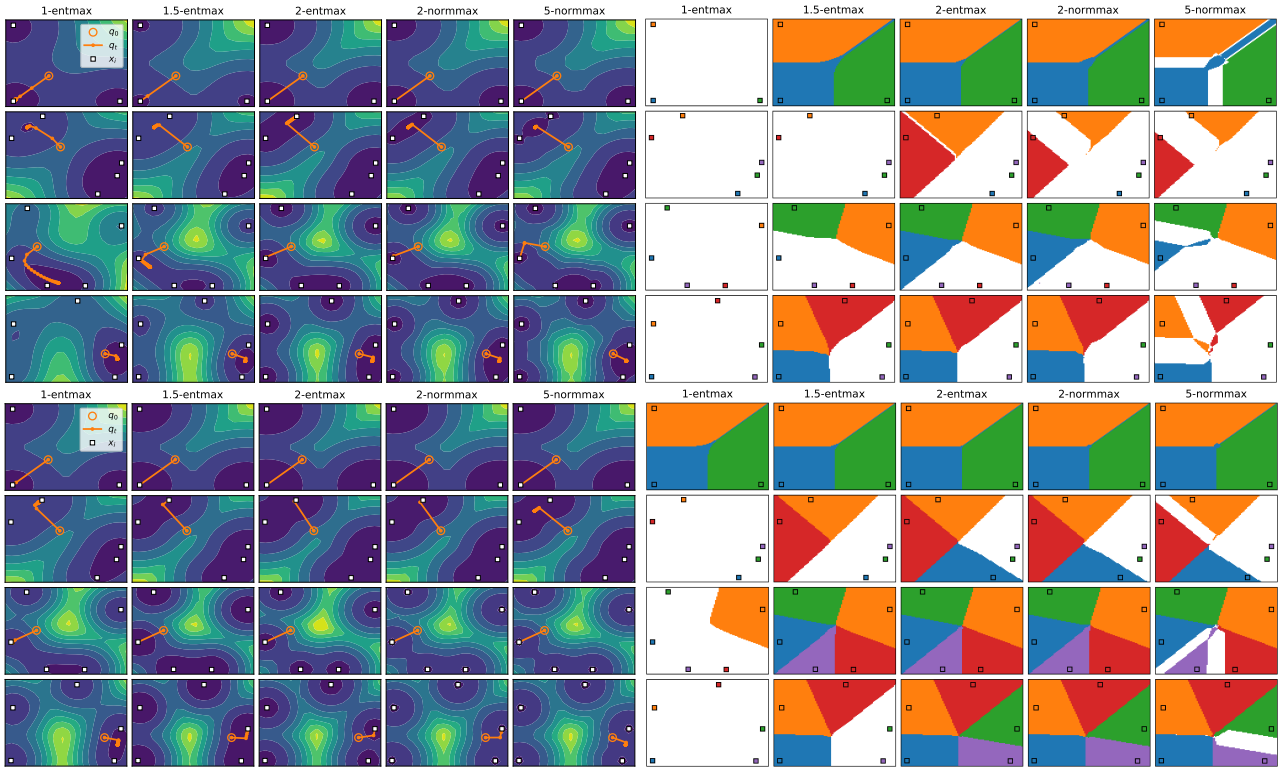

Figure 6: Top: Contours of the energy function and optimization trajectory of the CCCP iteration (left) alongside the attraction basins associated with each pattern (right) for $\beta = 4$. (Note that white sections do not converge to a single pattern but to a metastable state.) Bottom: Similar representation for $\beta = 10$.

and entmax becomes argmax, returning a one-hot vector. However, when $\alpha \to \infty$, the normmax regularizer becomes $\Omega_\infty^N(\boldsymbol{y}) = -1 + \max_i(y_i)$ and the transformation returns a uniform distribution with a sparse support, as shown in Figure 5, rightmost middle plot. On the other hand, when $\alpha \to 1$, entmax becomes softmax, but the normmax regularizer vanishes and normmax becomes argmax. SparseMAP with k-subsets, as β increases, tends to return a k-hot vector, where for $k = 1$ it corresponds to entmax with $\alpha = 2$ (sparsemax).

E.2. Hopfield dynamics and basins of attraction

In Figure 6, additional plots with varying β values are provided. Particularly, as β increases, the optimization trajectories exhibit a tendency to converge towards a single pattern, contrasting with the prevalence of metastable states observed for smaller β values. Moreover, the basins of attraction become increasingly colorful with higher β values, suggesting a convergence behaviour similar to what was previously described.

F. Experimental Details

F.1. MNIST K-MIL

For K-MIL, we created 4 datasets by grouping the MNIST examples into bags, for $K \in \{1, 2, 3, 5\}$. A bag is positive if it contains at least K targets, where the target is the number "9" (we chose "9" as it can be easily misunderstood with "7" or "4"). The embedding architecture is the same as Ilse et al. (2018), but instead of attention-based pooling, we use our α-entmax pooling, with $\alpha = 1$ mirroring the pooling method in (Ramsauer et al., 2021), and $\alpha = 2$ corresponding to the pooling in (Hu et al., 2023). Additionally, we incorporate α-normmax pooling and SparseMAP pooling with k-subsets. Further details of the K-MIL datasets are shown in Table 4.

We train the models for 5 different random seeds, where the first one is used for tuning the hyperparameters. The reported test accuracies represent the average across these seeds. We use 500 bags for testing and 500 bags for validation. The

Table 4: Dataset sample details for the MNIST K-MIL experiment. The size L_i of the ith bag is determined through $L_i = \max\{K, L'_i\}$ where $L'_i \sim \mathcal{N}(\mu, \sigma^2)$. The number of positive instances in a bag is uniformly sampled between K and L_i for positive bags and between 0 and $K-1$ for negative bags.

Dataset	μ	σ	Features	Pos. training bags	Neg. training bags
MNIST, $K=1$	10	1	28×28	1000	1000
MNIST, $K=2$	11	2	28×28	1000	1000
MNIST, $K=3$	12	3	28×28	1000	1000
MNIST, $K=5$	14	5	28×28	1000	1000

Table 5: Hyperparameter space for the MNIST MIL experiment. Hidden size is the dimension of keys and queries and γ is a parameter of the exponential learning rate scheduler (Li & Arora, 2020).

Parameter	Range
learning rate	$\{10^{-5}, 10^{-6}\}$
γ	$\{0.98, 0.96\}$
hidden size	$\{16, 64\}$
number of heads	$\{8, 16\}$
β	$\{0.25, 0.5, 1.0, 2.0, 4.0, 8.0\}$
bag dropout	$\{0.0, 0.75\}$

hyperparameters are tuned via grid search, where the grid space is shown in Table 5. We consider a dropout hyperparameter, commonly referred to as bag dropout, to the Hopfield matrix due to the risk of overfitting (as done by Ramsauer et al. (2021)). All models were trained for 50 epochs. We incorporated an early-stopping mechanism, with patience 5, that selects the optimal checkpoint based on performance on the validation set.

F.2. MIL benchmarks

The MIL benchmark datasets (Fox, Tiger and Elephant) comprise preprocessed and segmented color images sourced from the Corel dataset (Ilse et al., 2018). Each image is composed of distinct segments or blobs, each defined by descriptors such as color, texture, and shape. The datasets include 100 positive and 100 negative example images, with the negative ones randomly selected from a pool of photos featuring various other animals.

The $\mathrm{HopfieldPooling}$ layers (α-entmax; α-normmax; SparseMAP, k-subsets) take as input a collection of embedded instances, along with a trainable yet constant query. This query pattern is used for the purpose of averaging class-indicative instances, thereby facilitating the compression of bags of variable sizes into a consistent representation. This compression is important for effectively discriminating between different bags. To tune the model, a manual hyperparameter search was conducted on a validation set.

In our approach to tasks involving Elephant, Fox and Tiger, we followed a similar architecture as (Ramsauer et al., 2021):

1. The first two layers are fully connected linear embedding layers with ReLU activation.
2. The output of the second layer serves as the input for the $\mathrm{HopfieldPooling}$ layer, where the pooling operation is executed.
3. Subsequently, we employ a single layer as the final linear output layer for classification with a sigmoid as the classifier.

During the hyperparameter search, various configurations were tested, including different hidden layer widths and learning rates. Particular attention was given to the hyperparameters of the $\mathrm{HopfieldPooling}$ layers, such as the number of heads, head dimension, and the inverse temperature β. To avoid overfitting, bag dropout (dropout at the attention weights) was implemented as the chosen regularization technique. All hyperparameters tested are shown in Table 6. We trained for 50 epochs with early stopping with patience 5, using the Adam optimizer (Loshchilov & Hutter, 2017) with exponential learning rate decay. Model validation was conducted through a 10-fold nested cross-validation, repeated five times with different data splits where the first seed is used for hyperparameter tuning. The reported test ROC AUC scores represent the average across these repetitions.

Table 6: Hyperparameter space for the MIL benchmark experiments. Hidden size is the space in which keys and queries are associated and γ is a parameter of the exponential learning rate scheduler.

Parameter	**Range**
learning rate	$\{10^{-3}, 10^{-5}\}$
γ	$\{0.98 , 0.96\}$
embedding dimensions	$\{32 , 128\}$
embedding layers	$\{2\}$
hidden size	$\{32, 64\}$
number of heads	$\{12\}$
β	$\{0.1, 1, 10\}$
bag dropout	$\{0.0, 0.75\}$

Table 7: Text rationalization results. We report mean and min/max F_1 scores across five random seeds on test sets for all datasets but Beer, where we report MSE. All entries except SparseMAP are taken from (Guerreiro & Martins, 2021). We also report human rationale overlap (HRO) as F_1 score. We bold the best-performing rationalized model(s).

Method	Rationale	SST↑	AgNews↑	IMDB↑	Beer↓	Beer(HRO)↑
SFE	top-k	.76 (.71/.80)	.92 (.92/.92)	.84 (.72/.88)	.018 (.016/.020)	.19 (.13/.30)
	contiguous	.71 (.68/.75)	.86 (.85/.86)	.65 (.57/.73)	.020 (.019/.024)	35 (.18/.42)
SFE w/Baseline	top-k	.78 (.76/.80)	.92 (.92/.93)	.82 (.72/.88)	.019 (.017/.020)	.17 (.14/.19)
	contiguous	.70 (.64/.75)	.86 (.84/.86)	.76 (.73/.80)	.021 (.019/.025)	.41(.37/.42)
Gumbel	top-k	.70 (.67/.72)	.78 (.73/.84)	.74 (.71/.78)	.026 (.018/.041)	.27 (.14/.39)
	contiguous	.67 (.67/.68)	.77 (.74/.81)	.72 (.72/.73)	.043 (.040/.048)	.42 (.41/.42)
HardKuma	-	.80 (.80/.81)	.90 (.87/.88)	.87 (.90/.91)	.019 (.016/.020)	.37 (.00/.90)
Sparse Attention	sparsemax	**.82** (.81/.83)	**.93** (.93/.93)	.89 (.89/.90)	.019 (.016/.021)	48 (.41/.55)
	fusedmax	.81 (.81/.82)	.92 (.91/.92)	.88 (.87/.89)	.018 (.017/.019)	39 (.29/.53)
SPECTRA	sequential k-subsets	.80 (.79/.81)	.92 (.92/.93)	**.90** (.89/.90)	**.017** (.016/.019)	.61 (.56/.68)
SparseMAP	k-subsets	.81 (.81/.82)	**.93** (.92/.93)	**.90** (.90/.90)	**.017** (.017/.018)	.42 (.29/.62)
	sequential k-subsets	.81 (.80/.83)	**.93** (.93/.93)	**.90** (.90/.90)	.020 (.018/.021)	**.63** (.49/.70)

F.3. Text Rationalization

We use the same hyperparameters reported by Guerreiro & Martins (2021). We used a head dimension of 200, to match the dimensions of the encoder vectors (the size of the projection matrices associated to the static query and keys) and a head dropout of 0.5 (applied to the output of the Hopfield layer). We used a single attention head to better match the SPECTRA model. Aditionally we use a transition score of 0.001 and a train temperature of 0.1.

We present in Table 7 an extended version of Table 3 with additional baselines, corresponding to prior work in text rationalization. We show additional examples of rationales extracted by our models in Figure 7.

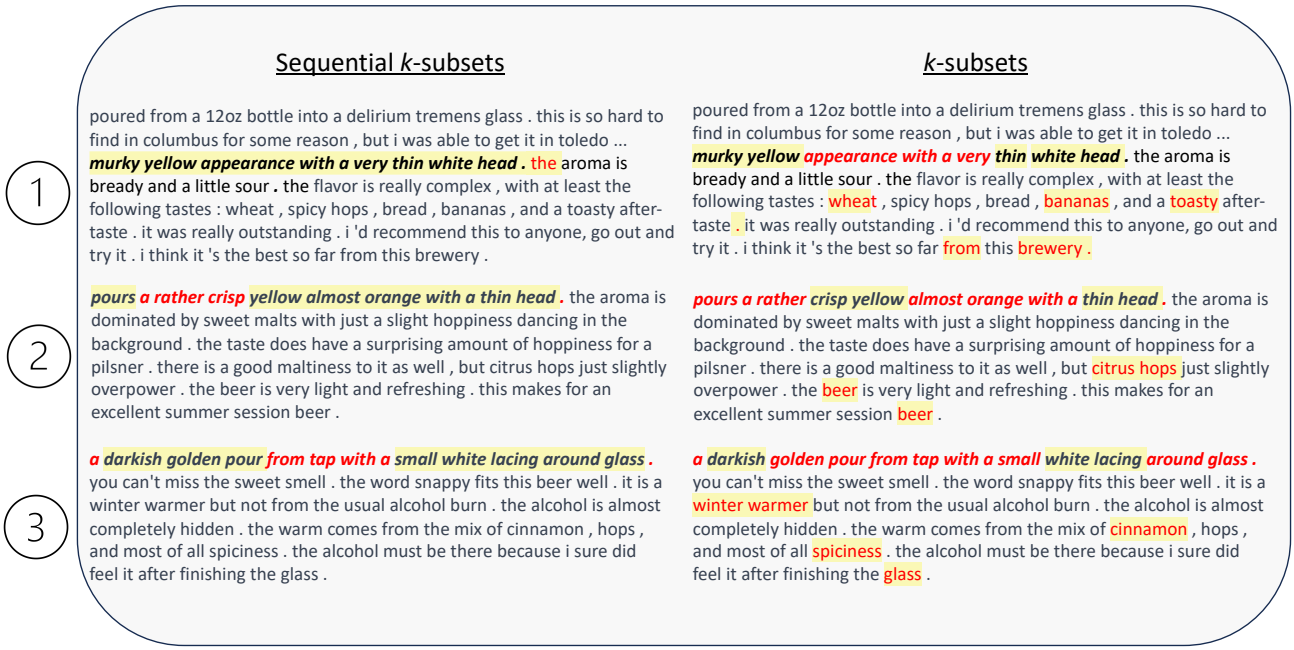

Figure 7: Examples of human rationale overlap for the aspect "appearance". The yellow highlight indicates the model's rationale, while ***italicized and bold font*** represents the human rationale. Red font identifies mismatches with human annotations. SparseMAP with sequential k-subsets prefers more contiguous rationales, which better match humans.