Uniform Memory Retrieval with Larger Capacity for Modern Hopfield Models

Dennis Wu*† **Jerry Yao-Chieh Hu***† **Teng-Yun Hsiao** ♮ **Han Liu**†

† Northwestern University, Evanston, IL 60208 USA; ♮ National Taiwan University, Taipei 10617, Taiwan

Northwestern University

National Taiwan University

Summary

- We propose to use a learnable linear kernel as the similarity measure in modern Hopfield models, resulting in kernelized memory Hopfield energy.
- We propose a two-stage retrieval dynamics termed `U-Hop`. The first stage maximizes the separation between memories in kernel space by optimizing the linear kernel. The second stage performs energy minimization with kernel induced retrieval dynamics.
- Empirically, `U-Hop` improves memory retrieval outcomes by a large margin comparing to other baselines. When applied to deep learning scenarios, `U-Hop` significantly improves model's memorization capacity, generalization and convergence speed.

Kernelized Memory Hopfield Model

Let $\mathbf{x} \in \mathbb{R}^d$ be the input query, and $\mathbf{\Xi} := \{\boldsymbol{\xi}\}_{\mu=1}^M \in \mathbb{R}^{M\times d}$ be the memory set.

Definition 1 (α-EntMax) Let $\mathbf{z}, \mathbf{p} \in \mathbb{R}^M$, and $\Delta^M := \{\mathbf{p} \in \mathbb{R}_+^M \mid \sum_\mu^M p_\mu = 1\}$ be the $(M-1)$-dimensional unit simplex. Let $\Psi_\alpha(\mathbf{p})$ be the Tsallis α-entropy The α-EntMax is defined as

$$\alpha\text{-}\mathit{EntMax}(\mathbf{z}) := \underset{\mathbf{p}\in\Delta^M}{\mathit{ArgMax}}[\langle \mathbf{p}, \mathbf{z}\rangle - \Psi_\alpha(\mathbf{p})].$$

Let $\mathcal{K}(\cdot,\cdot) := \langle \Phi(\cdot), \Phi(\cdot)\rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, where $\Phi : \mathbb{R}^d \to \mathbb{R}^{D_\Phi}$ and $D_\Phi \gg d$. Let feature map Φ be $\Phi(\mathbf{u}) := \mathbf{W}\mathbf{u}$ with $\mathbf{W} \in \mathbb{R}^{D_\Phi \times d}$ for any $\mathbf{u} \in \mathbb{R}^d$.

Assumption 1 $\boldsymbol{W} \in \mathbb{R}^{D_\Phi \times d}$ with $D_\Phi >> d$ is full rank.

Definition 2 (Kernelized Memory Hopfield Energy) Let $\mathbf{x} \in \mathbb{R}^d$ be the input query, $\mathbf{\Xi} \in \mathbb{R}^{d\times M}$ be the stored memory set.

$$E_\mathcal{K}(\mathbf{x}) = \frac{1}{2}\mathcal{K}(\mathbf{x},\mathbf{x}) - \Psi_\alpha^\star\left(\beta, \mathcal{K}(\mathbf{\Xi}^\top \mathbf{x})\right),$$

where $\Psi_\alpha^\star$ is the convex conjugate of the Tsallis entropic regularizer.

Theorem 1 (Kernelized Memory Hopfield Retrieval Dynamics) With Assumption 1, the energy function $E(\mathbf{x})$ was monotonically decreased by the following retrieval dynamics:

$$\mathcal{T}_\mathcal{K}(\mathbf{x}) = \mathbf{\Xi} \cdot \alpha\text{-EntMax}\left(\beta \cdot \mathcal{K}(\mathbf{\Xi}, \mathbf{x})\right).$$

- The assumption of $\mathbf{W}$ being full column rank is necessary for us to be able to project memories into feature space and back.
- The full-column rank assumption is also critical for kernelized memory Hopfield models to preserve the defining properties of modern Hopfield models.
- The Kernelized Memory Hopfield construction is compatible for various of existing modern Hopfield models, such as modern Hopfield , sparse modern Hopfield, etc.

Memory Separation

Definition 3 (Pattern Stored and Retrieved) For all $\mu \in [M]$, let $R_\Phi := \frac{1}{2}\,\mathrm{Min}_{\nu\neq\mu;\nu,\mu\in[M]} \|\Phi(\boldsymbol{\xi}_\mu) - \Phi(\boldsymbol{\xi}_\nu)\|$ be the finite radius of each (kernelized) sphere $\mathcal{S}_{\Phi,\mu}$ centered at (kernelized) memory pattern $\Phi(\boldsymbol{\xi}_\mu)$. We say $\boldsymbol{\xi}_\mu$ is *stored* if there exists a generalized fixed point of $\mathcal{T}_\mathcal{K}$, such that $\Phi(\mathbf{x}_\mu^\star) \in \mathcal{S}_{\Phi,\mu}$, to which all limit points $\Phi(\mathbf{x}) \in \mathcal{S}_{\Phi,\mu}$ converge to, and $\mathcal{S}_{\Phi,\mu} \cap \mathcal{S}_{\Phi,\nu} = \emptyset$ for $\nu \neq \mu$. We say $\boldsymbol{\xi}_\mu$ is ϵ-*retrieved* by $\mathcal{T}_\mathcal{K}$ with $\mathbf{x}$ for an error ϵ.

Lemma 1 (Retrieval Error Bound of $\mathcal{T}_\mathcal{K}$) Let $\Delta_\mu^\Phi := \mathcal{K}(\boldsymbol{\xi}_\mu), \boldsymbol{\xi}_\mu) - \mathrm{Max}_{\nu\in[M],\nu\neq\mu}\,\mathcal{K}(\boldsymbol{\xi}_\nu, \boldsymbol{\xi}_\mu)$ be the separation between a memory pattern $\boldsymbol{\xi}_\mu$ from all other memories in the feature space. Assuming patterns are normalized in feature space, we have

$$\|\mathcal{T}_\mathcal{K}(\mathbf{x}) - \boldsymbol{\xi}_\mu\| \leq 2(M-1)e^{-\beta(\Delta_\mu^\Phi - 2R_\Phi)}.$$

In comparison, the retrieval error bound of the modern Hopfield models is

$$\|\mathcal{T}_{\mathsf{MHM}}(\mathbf{x}) - \boldsymbol{\xi}_\mu\| \leq 2(M-1)e^{-\beta(\Delta_\mu - 2R)},$$

where $\Delta_\mu := \langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\mu\rangle - \mathrm{Max}_{\nu\in[M],\nu\neq\mu}\,\langle \boldsymbol{\xi}_\nu, \boldsymbol{\xi}_\mu\rangle$.

A critical aspect of our kernelized memory Hopfield models is we successfully relax the dependency on Δ_μ to Δ_μ^Φ, which can be optimized by searching for a better Φ.

`U-Hop`: Two Stage Memory Retrieval

To search for a better Φ with larger separation, we propose to optimize Φ under the average separation loss:

Definition 4 (Average Separation Loss) Given a stored memory set $\mathbf{\Xi}$, and a feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^{D_\Phi}$, the separation loss of the function Φ is $\mathcal{L}_\Phi(\mathbf{\Xi}; t) := \log \underset{\mathbf{u},\mathbf{v}\sim\mathbf{\Xi}}{\mathbb{E}}[\mathcal{G}_t(\Phi(\mathbf{u}), \Phi(\mathbf{v}))],\ t > 0$.

- Minimization of $\mathcal{L}_\Phi$ leads to an on-average dissimilarity among kernelized memory patterns, i.e., $\{\Phi(\xi_\mu)\}_{\mu\in[M]}$.
- $\mathcal{L}_\Phi$ is convex by design and hence exists an optimizer $\mathbf{W}^\star$ that maximizes the average distance between all possible memory pairs.

Algorithm 1 `U-Hop`: Two-Stage Memory Retrieval

Input: Separation (Stage I) iterations N, Energy (Stage II) iteration T, feature map $\Phi(\mathbf{x}) := \mathbf{W}\mathbf{x}$, memory set $\mathbf{\Xi}$, query $\mathbf{x}$, retrieval dynamics $\mathcal{T}$, learning rate $\gamma \leq 1/G$ where G is the Lipschitz constant of $\mathcal{L}_\Phi(\mathbf{\Xi})$

Output: $\mathbf{x}$

for $i = 1, \ldots N$ **do**
  $\mathbf{W} \leftarrow \mathbf{W} - \gamma \cdot \nabla_\mathbf{W}\mathcal{L}_\Phi(\mathbf{\Xi})$. // Stage I
end for
Normalize the rows of $\mathbf{W}$
$\mathbf{x}^0 \leftarrow \mathbf{x}$
for $t = 1, \ldots T$ **do**
  $\mathbf{x} \leftarrow \mathcal{T}_\mathcal{K}(\mathbf{x})$ using Theorem 2.1 // Stage II
end for
return $\mathbf{x}$

Exact Memory Retrieval with `U-Hop`

With controllable separation, we are able to obtain the conditions of exact retrieval under $\mathcal{T}_\mathcal{K}$.

Theorem 2 (Exact Memory Retrieval) Let $\mathcal{T}_{\mathsf{sparse}}$ be $\mathcal{T}_\mathcal{K}$ from retrieval dynamics with $\alpha > 1$. Let $\mathcal{K}$ be a real-valued kernel with feature map Φ. Let $t > 0, \beta > 0$. Supposed the query $\mathbf{x} \in \mathcal{S}_{\Phi,\mu}$, $\Phi(\xi_\mu)$ is the fixed point of $\mathcal{T}_{\mathsf{sparse}}$ if the following condition is satisfied:

$$\ell_\Phi(\xi_\mu, \xi_\mu) - \max_{\nu,\nu\neq\mu} \ell_\Phi(\xi_\nu, \xi_\mu) \leq -\frac{2t}{\beta(\alpha-1)}.$$

Further, let $L > 0$ be the Lipschitz constant of Φ. Following the above result, $\mathcal{T}_\mathcal{K}$ achieves exact memory retrieval if

$$\underset{\nu\in[M],\nu\neq\mu}{\mathrm{Min}} \|\xi_\mu - \xi_\nu\| \geq \sqrt{\frac{2}{L^2\beta(\alpha-1)}}.$$

Experimental Studies

Memory Retrieval Error Comparison:

- All four plots show `U-Hop` retrieved patterns with significantly less error compared to all existing baselines across all sizes of memory and noise levels.

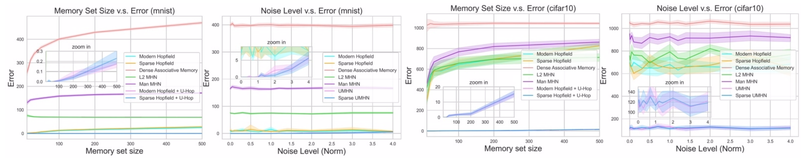

Image Classification Task:

- The result demonstrates with `U-Hop`, models are able to consistently memorize more samples in the training data, and further obtain generalization improvement.
- For interpretation on maximal training accuracy, please refer to Theorem 3.1 in the original paper.

Models	CIFAR10		CIFAR100		Tiny ImageNet	
	Max Train Acc.	Test Acc.	Max Train Acc.	Test Acc.	Max Train Acc.	Test Acc.
MHM	56.0%	52.2%	32.3%	26.3%	48.9%	12.2%
MHM + U-Hop	**64.6%**	**55.2%**	**44.1%**	**28.7%**	**61.4%**	**12.7%**
Sparse MHM	55.9%	52.0%	49.6%	26.0%	17.2%	12.3%
Sparse MHM + U-Hop	**66.4%**	**55.4%**	**45.4%**	**29.0%**	**60.6%**	**12.5%**

Model Convergence Comparison on CIFAR100:

- Left to right: Training Accuracy, Test Accuracy, Training and Test Loss.
- For generalization power and convergence speed, models with `U-Hop` outperform other baselines by a large margin.