AAAI2025 Association for the Advancement of Artificial Intelligence

Recognition and Learning Algorithm Laboratory

Rectified Lagrangian for Out-of-Distribution Detection in Modern Hopfield Networks

Ryo Moriai[1], Nakamasa Inoue[1], Masayuki Tanaka[1], Rei Kawakami[1], Satoshi Ikehata[1,3], Ikuro Sato[1,2]

[1]Institute of Science Tokyo, Japan, [2]Denso IT Laboratory, Japan, [3] National Institute of Informatics, Japan

DENSO IT LAB

Institute of SCIENCE TOKYO

1. Introduction

Background- Modern Hopfield Networks (MHNs)

MHNs are energy-based models that can retrieve a semantically relevant data for a given query data.

Issue of MHN

For an out-of-distribution (OOD) query, MHN inevitably associates an inappropriate in-distribution (ID) data.

Contributions

- To address this, we propose **Rectified Lagrangian (RegLag)** that explicitly incorporates OOD queries with a specially designed attractor in the dynamical system of MHNs.
- RecLag-based MHNs yield higher OOD detection performance over existing energy-based methods on average using nine image datasets.

2. Large-Memory MHNs

[D. Krotov+, ICLR2021; M. Widrich+, NeurIPS2021]

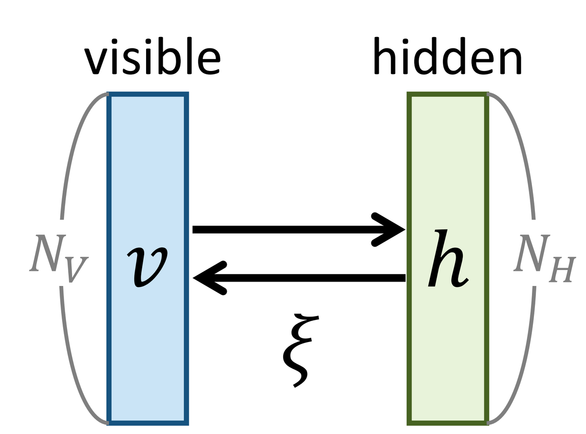
General Formulation of Energy Function

$$E_\xi(v,h) =$$

$$v^\top \frac{\partial L_V(v)}{\partial v} - L_V(v)$$ Energy associated with visible layer

$$+h^\top \frac{\partial L_H(h)}{\partial h} - L_H(h)$$ Energy associated with hidden layer

$$-\frac{\partial L_H(h)}{\partial h}^\top \xi \frac{\partial L_V(v)}{\partial v}$$ Interaction energy between visible & hidden layers



Large-Memory Lagrangian Functions

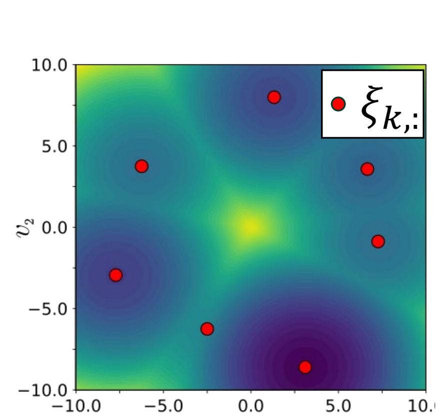
$$L_V = \frac{1}{2}\sum_{i=1}^{N_V} v_i^2 \qquad L_H = \frac{1}{\beta}\log\sum_{k=1}^{N_H} e^{\beta h_k}$$

$(\beta > 0)$

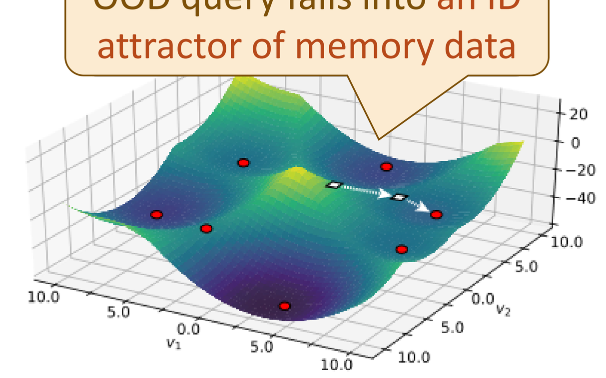
Proven that log-sum-exp assures exponentially-ordered memory capacity.

Discrete Update Rule

$v \leftarrow \xi^\top \mathrm{softmax}(\beta \xi v)$ Converges to an **attractor**: $v \to \xi_{k,:}$ (under $0 < \beta \ll 1$)



Energy landscape with attractors



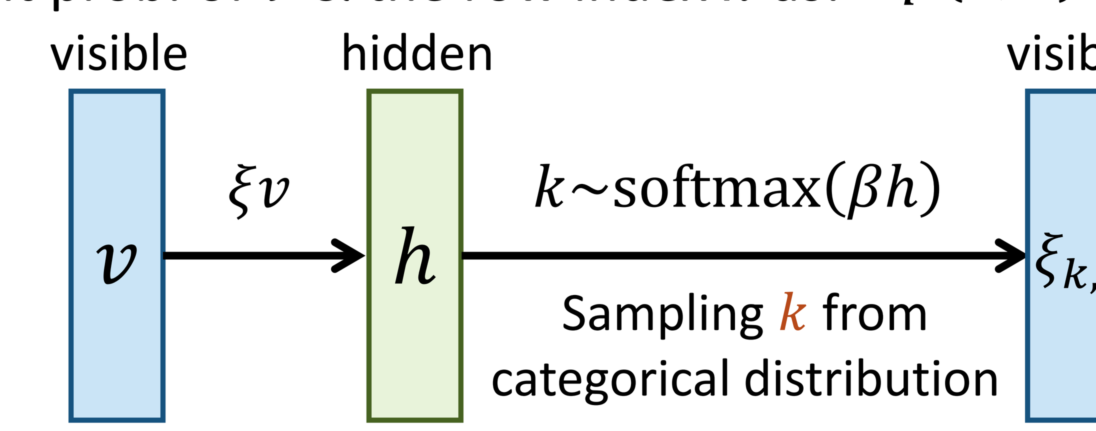
Issue: An OOD query is associated with an irrelevant attractor.

3. Proposed Method: RecLag

Probability-Density Aware MHNs

We *optimize* ξ to express joint prob. of v & the row index k as: $p(v,k) \propto e^{\beta \xi^*_{k,:} v} \propto \mathrm{softmax}(\beta \xi^* v)_k$

Step 1) Unrolling RNN



The attractor dynamics is virtually acquired when the input is reconstructed.

Step 2) Optimizing probabilistic AE $\xi^* = \arg\max_\xi \prod_{v_j \in \mathcal{D}} \prod_{\tilde{v} \sim p(v)} \mathcal{N}(\tilde{v} | v_j; I_{M \times M})$

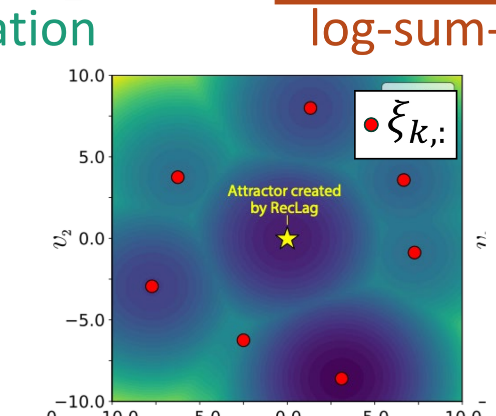
The probability density for v is given by: $\log p(v) = \underbrace{\log \sum_{k=1}^{N_H} e^{\beta \xi_k v}}_{\text{hidden-layer Lagrangian } L_H} + \text{const}$

Now, log-sum-exp is connected to $p(v)$.

Rectifed Lagrangian Containing a Trivial Attractor

We replace the hidden-layer Lagrangian with:

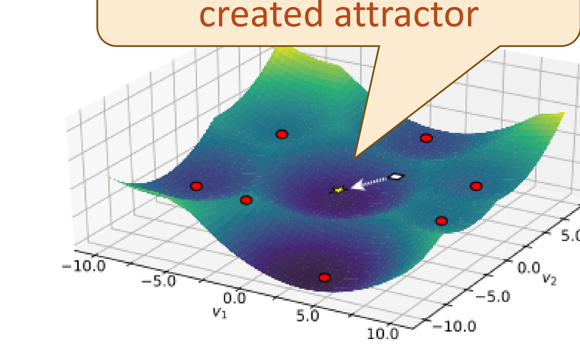
$$L_H^{\mathrm{Rec}}(h) = \underbrace{\max}_{\text{rectification}}\left[\frac{1}{\beta}\log\gamma, \underbrace{\frac{1}{\beta}\log\sum_{k=1}^{N_H} e^{\beta h_k}}_{\text{log-sum-exp}}\right] \qquad \gamma > 0$$



Energy landscape with attractors

Modified Update Rule

$$v \leftarrow \begin{cases} \xi^\top \mathrm{softmax}(\beta \xi v), & L_H(h) \geq \beta^{-1}\log\gamma \\ \underbrace{0,}_{\text{attractor absorbing OOD queries}} & L_H(h) < \beta^{-1}\log\gamma \end{cases}$$



An OOD query is associated with the created attractor.

4. Evaluation

Outline

OOD detection performance is evaluated against 5 baselines using 9 datasets.

Result

- RecLag-based MHNs yield higher OOD detection performance over baselines on average.
- RecLag-based MHNs also showed the highest accuracy in 23 out of 27 individual experimental settings.

[D. Hendrycks+, ICLR2017; W. Liu+, NeurIPS2020; Y. Sun+, NeurIPS2021; J Zhang+, ICLR2023]

Table: OOD detection performance as FPR95(%) ↓

	Method	SVHN	LSUN-C	LSUN-R	iSUN	Places	DTD	TIN	SUN	iNaturalist	Average
ResNet18	MSP	76.34	27.52	36.54	34.84	20.55	30.65	45.82	22.89	12.62	34.19
	Energy	56.05	8.10	11.60	9.10	3.18	16.98	25.47	3.27	3.47	15.25
	ReAct	59.47	7.57	12.52	10.13	2.93	16.86	27.61	3.27	3.80	16.02
	MHE	17.59	9.20	7.68	4.74	0.33	8.96	15.86	**0.00**	2.35	7.41
	SHE	**17.45**	9.22	7.69	4.77	0.33	8.99	15.84	**0.00**	2.38	7.41
	RecLag	18.12 ± 2.02	**6.40 ± 0.25**	**4.60 ± 0.12**	**2.67 ± 0.47**	**0.28 ± 0.02**	**6.82 ± 0.13**	**12.09 ± 0.25**	**0.00 ±0.00**	**1.68 ± 0.04**	**5.85 ± 0.24**
ResNet34	MSP	59.86	28.26	32.06	31.69	33.61	43.28	45.56	32.43	32.95	37.74
	Energy	30.51	6.84	9.43	8.47	9.32	23.74	25.16	8.99	10.86	14.81
	ReAct	45.86	14.37	14.09	13.28	15.83	29.73	31.60	15.53	11.98	21.36
	MHE	6.20	6.17	4.40	2.94	2.34	14.32	15.86	0.54	4.91	6.41
	SHE	6.14	6.20	4.45	3.01	2.36	14.32	15.93	0.54	4.92	6.43
	RecLag	**5.19 ± 0.24**	**5.60 ± 0.07**	**2.85 ± 0.05**	**2.11 ± 0.05**	**2.31 ± 0.03**	**12.04 ± 0.07**	**11.71 ± 0.23**	**0.33 ± 0.11**	**4.14 ± 0.08**	**5.14 ± 0.08**
WRN40-2	MSP	41.52	44.43	38.47	39.70	33.84	35.80	51.52	34.88	27.69	38.65
	Energy	15.35	17.77	14.98	17.45	10.58	19.71	36.75	9.54	8.95	16.79
	ReAct	18.83	19.93	18.25	20.68	11.98	21.67	42.02	11.44	13.26	19.78
	MHE	5.40	14.60	12.03	11.48	2.90	10.99	27.28	**0.82**	1.83	9.70
	SHE	**5.25**	14.39	13.18	12.39	2.83	10.98	28.35	**0.82**	1.84	10.00
	RecLag	5.75 ± 0.12	**7.37 ± 0.18**	**8.44 ± 0.17**	**8.01 ± 0.15**	**2.63 ± 0.05**	**9.75 ± 0.10**	**22.62 ± 0.34**	1.06 ± 0.09	**1.67 ± 0.05**	**7.47 ± 0.85**

This work was supported by JSPS KAKENHI Grant Number JP22H03642 and DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Science Tokyo).

AAAI2025

Association for the Advancement of Artificial Intelligence

Recognition and Learning Algorithm Laboratory

Rectified Lagrangian for Out-of-Distribution Detection in Modern Hopfield Networks

Ryo Moriai[1], Nakamasa Inoue[1], Masayuki Tanaka[1], Rei Kawakami[1], Satoshi Ikehata[1,3], Ikuro Sato[1,2]

[1]Tokyo Institute of Technology, Japan, [2]Denso IT Laboratory, Japan, [3] National Institute of Informatics, Japan

DENSO IT LAB

Institute of SCIENCE TOKYO

Introduction

Modern Hopfield Networks (MHNs)

ergy-based models that can retrieve a relevant data for a given query data.

N

-distribution (OOD) query, MHN does n-distribution (ID) data, which is totally

s

ss this, we propose Rectified Lagrangian a new Lagrangian that explicitly ates OOD queries with a specially designed in the dynamical system of MHNs. ag-based MHNs demonstrated higher OOD n performance over existing energy-based ection methods on average using nine tasets.

2. Modern Hopfield Network (MHN) [D. Krotov+ 2021, Widrich et al. 2021]

Formulation

$v \in \mathbb{R}^{N_V}$: visible layer neuron $h \in \mathbb{R}^{N_H}$: hidden layer neuron,

β: real hyperparameter,

$\Xi \in \mathbb{R}^{N_H \times N_V}$: matrix parameter,

ξ_μ: μ-th column vector component of a matrix Ξ

Energy

$$E = \sum_{i=1}^{N_V} v_i\, g_i(v) - L_V(v) + \sum_{\mu=1}^{N_V} h_\mu f_\mu(h) - L_V(v) - f(h)^\top \Xi g(v)$$

Activation functions

$$f(h) = \frac{\partial L_H(h)}{\partial h} \qquad g(v) = \frac{\partial L_V(v)}{\partial v}$$

Hidden-layer Lagrangian

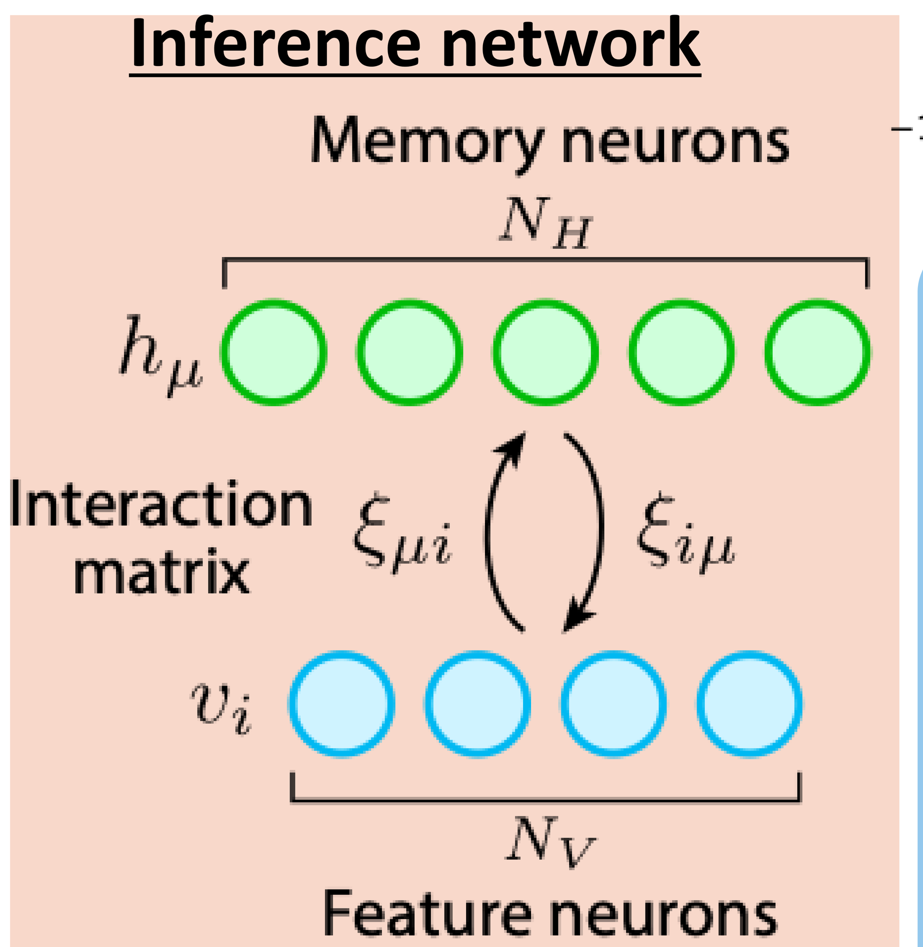
$$L_h = \frac{1}{\beta}\log\left(\sum_{\mu=1}^{N_H} e^{\beta h_\mu}\right)$$

Visible Lagrangian

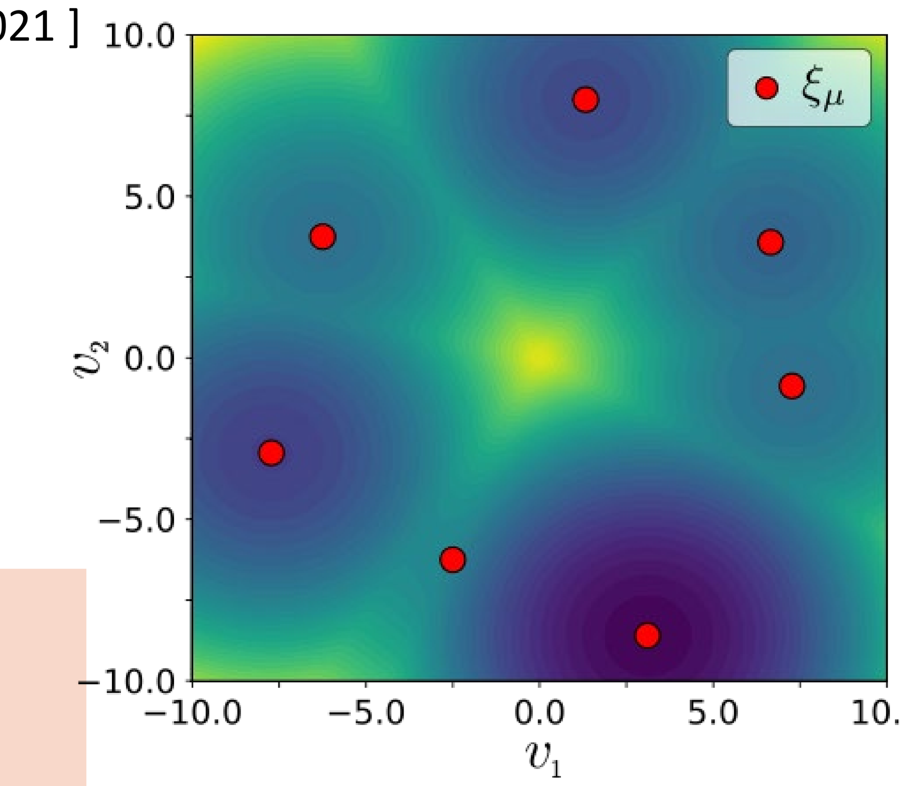
$$L_v = \frac{1}{2}\sum_{i=1}^{N_V} v_i^2$$

Discrete Update rule

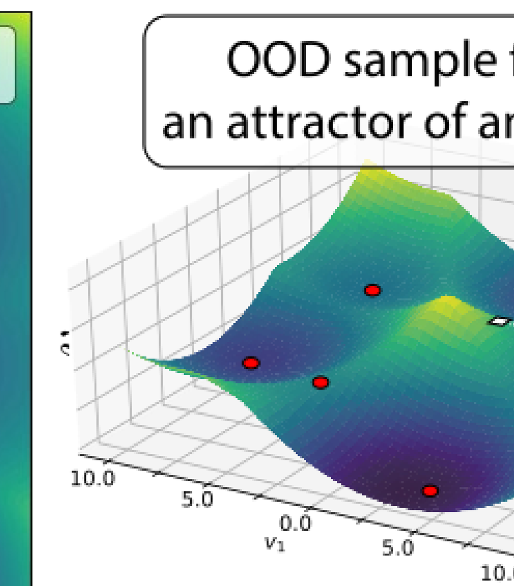
$$v^{k+1} = \Xi^\top \mathrm{softmax}(\beta \Xi v^k)$$



Energy Map



OOD sample



Proposed Method

ame energy except the hidden-gian is replaced by following

ayer Lagrangian (RecLag)

$$\left(\frac{1}{\beta}\log\left(\frac{1}{\gamma}\sum_{\mu=1}^{N_H} e^{\beta h_\mu}\right), 0\right)$$

$$\left(G(v^k)\right)\Xi^\top \mathrm{softmax}(\beta\Xi\, v^k)$$

$$\log\left(\frac{1}{\gamma}\sum_{\mu=1}^{N_H} e^{\beta \xi_\mu v}\right)$$

$(x \geq 0)$
$(x < 0)$

$\gamma \in \mathbb{R}_{\geq 0}$: Real hyperparameter

Probability-based learning

W is determined through optimization to satisfy the following probability equation.

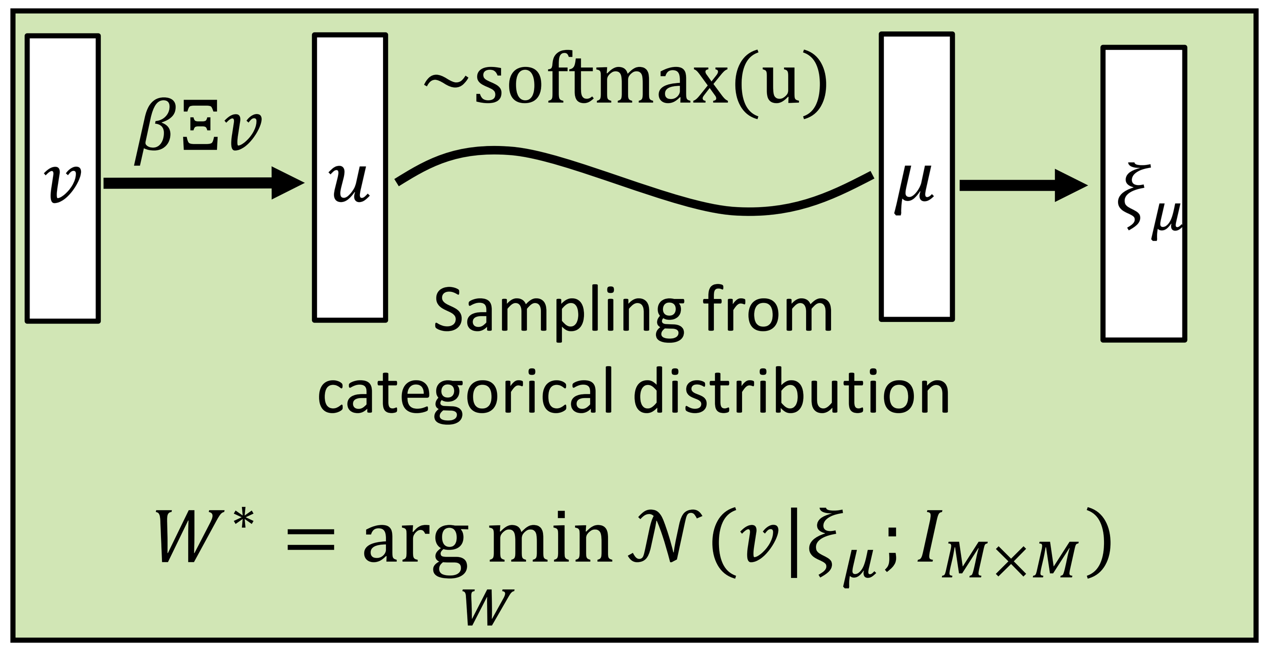
$$p_H(\mu, v) = \frac{1}{Z} e^{\beta \xi_\mu v} \propto f(h)_\mu$$

Each neuron in the hidden layer corresponds to the probability of the μ-th memory

$$\log(p(v)) - \log\left(\sum_{\mu=1}^{N_H} e^{\beta \xi_\mu v}\right) = \text{const}$$

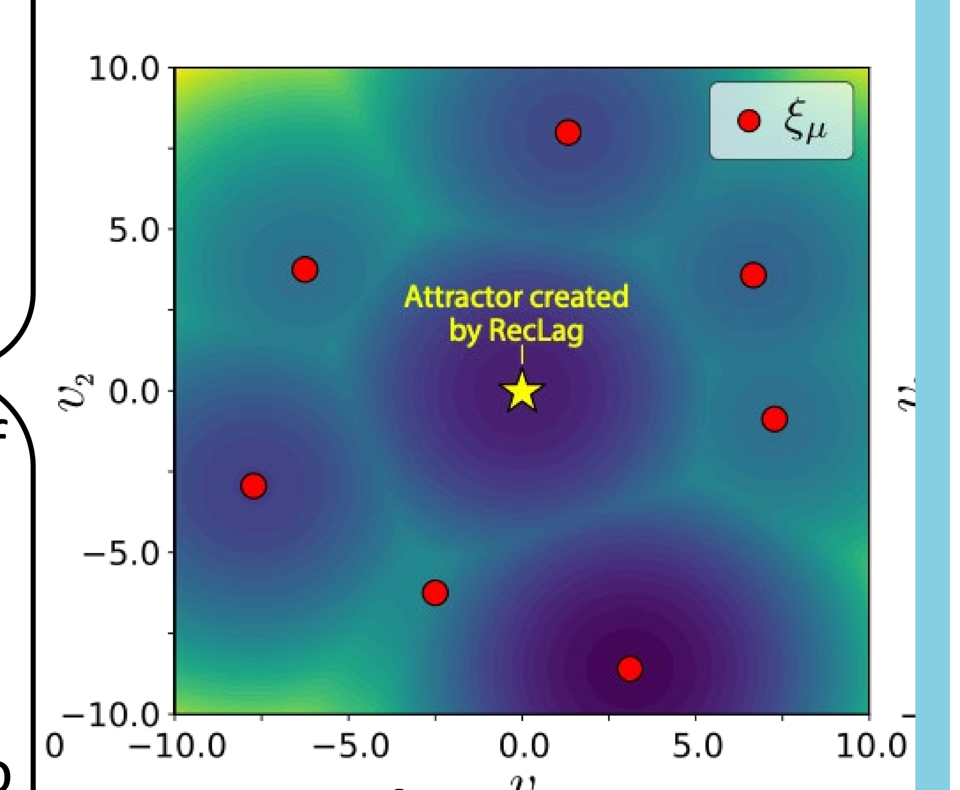
The Lagrangian of the hidden layer corresponds to the probability density relative to the memory's parent distribution.

$Z \in \mathbb{R}$: normalized coefficient

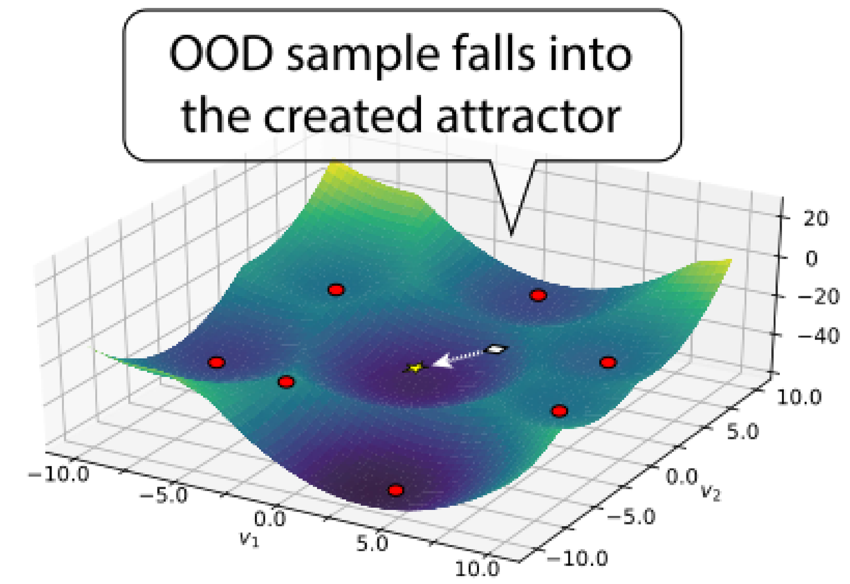


$\mathcal{N}$: multivariate normal distribution $I_{M\times M}$: M-dimensional identity matrix

Energy Map



OOD sample



4. Experiment

	Method	SVHN	LSUN-C	LSUN-R	iSUN	Places	DTD	TIN	SUN
ResNet18	MSP	76.34	27.52	36.54	34.84	20.55	30.65	45.82	22.8
	Energy	56.05	8.10	11.60	9.10	3.18	16.98	25.47	3.2
	ReAct	59.47	7.57	12.52	10.13	2.93	16.86	27.61	3.2
	MHE	17.59	9.20	7.68	4.74	0.33	8.96	15.86	**0.0**
	SHE	**17.45**	9.22	7.69	4.77	0.33	8.99	15.84	**0.0**
	RecLag	18.12 ± 2.02	**6.40 ± 0.25**	**4.60 ± 0.12**	**2.67 ± 0.47**	**0.28 ± 0.02**	**6.82 ± 0.13**	**12.09 ± 0.25**	**0.0 ±0.0**
ResNet34	MSP	59.86	28.26	32.06	31.69	33.61	43.28	45.56	32.4
	Energy	30.51	6.84	9.43	8.47	9.32	23.74	25.16	8.9
	ReAct	45.86	14.37	14.09	13.28	15.83	29.73	31.60	15.5
	MHE	6.20	6.17	4.40	2.94	2.34	14.32	15.86	0.5
	SHE	6.14	6.20	4.45	3.01	2.36	14.32	15.93	0.5
	RecLag	**5.19 ± 0.24**	**5.60 ± 0.07**	**2.85 ± 0.05**	**2.11 ± 0.05**	**2.31 ± 0.03**	**12.04 ± 0.07**	**11.71 ± 0.23**	**0.3 ± 0.1**
WRN40-2	MSP	41.52	44.43	38.47	39.70	33.84	35.80	51.52	34.8
	Energy	15.35	17.77	14.98	17.45	10.58	19.71	36.75	9.5
	ReAct	18.83	19.93	18.25	20.68	11.98	21.67	42.02	11.4
	MHE	5.40	14.60	12.03	11.48	2.90	10.99	27.28	**0.8**
	SHE	**5.25**	14.39	13.18	12.39	2.83	10.98	28.35	**0.8**
	RecLag	5.75 ± 0.12	**7.37 ± 0.18**	**8.44 ± 0.17**	**8.01 ± 0.15**	**2.63 ± 0.05**	**9.75 ± 0.10**	**22.62 ± 0.34**	1.0 ± 0.0

Objective

To evaluate OOD detection performance for RecLag.

Baselines

The FPR95 of the five OOD detection methods:{MSP,Energy,React,SHE,HE}

Settings

Features of a CIFAR-10 pretrained network a OOD detection. In-distribution data: CIFAR-1 OOD data: {SVHN, LSUN_C, LSUN_R, iSUN, Pl Tiny Image Net, SUN, iNaturalist}.

For the proposed RecLag the trimmed mean standard deviations (following ± symbols) o with the largest and the smallest ones being are reported.

This work was supported by JSPS KAKENHI Grant Number JP22H03642 and DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Science Tokyo).

en out-of-distribution (OOD) query, MHN does retrieve a memory data, totally irrelevant.

butions

dress this, we propose the rectified Lagrangian (RegLag), a new angian for memory neurons that explicitly incorporates an attractor for D samples in the dynamical system of MHNs.

demonstrate outperformance RecLag-based MHNs over energy-based D detection methods, including state-of-the-art Hopfield Energy, on nine ge datasets.

. Modern Hopfield Network (MHN)

lation [D. Krotov+ 2021, Widrich et al. 2021]

y

$$= \sum_{i=1}^{N_V} v_i\, g_i(v) - L_V(v) + \sum_{\mu=1}^{N_V} h_\mu f_\mu(h) - L_V(v) - f(h)^\top \Xi g(v)$$

tion functions

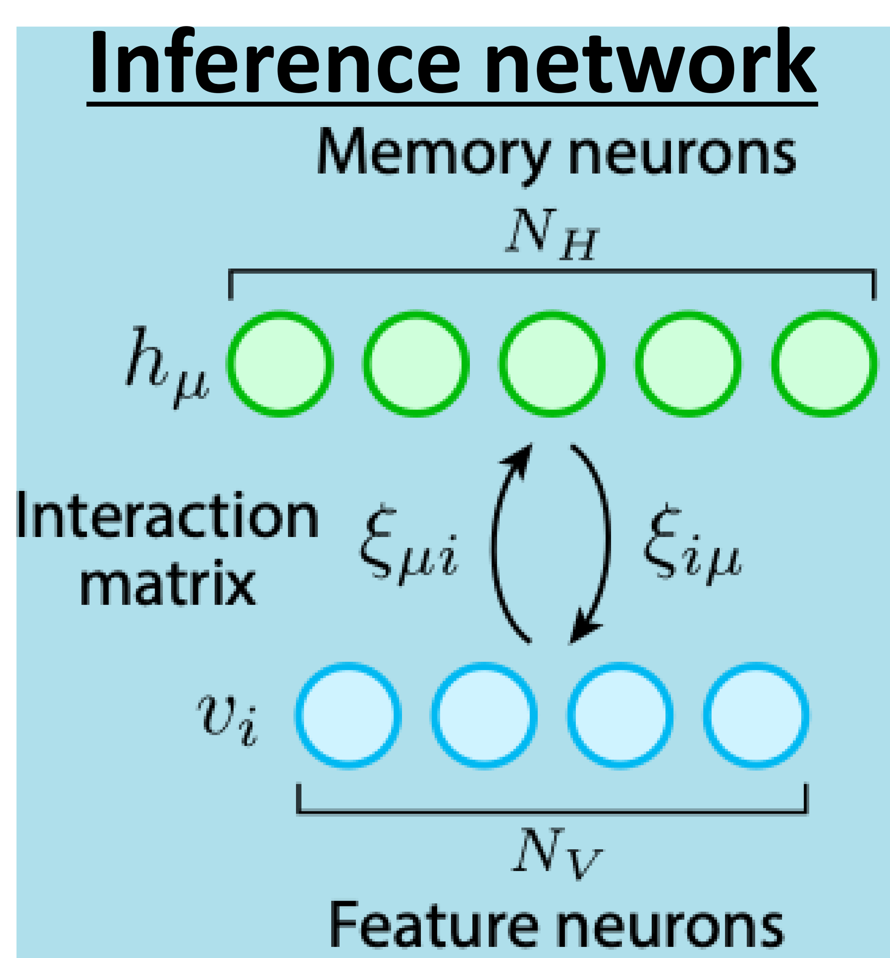
$$= \frac{\partial L_H(h)}{\partial h} \qquad g(v) = \frac{\partial L_V(v)}{\partial v}$$

n-layer Lagrangian

$$\frac{1}{\beta}\log\left(\sum_{\mu=1}^{N_H} e^{\beta h_\mu}\right)$$

Visible Lagrangian

$$L_v = \frac{1}{2}\sum_{i=1}^{N_V} v_i^2$$

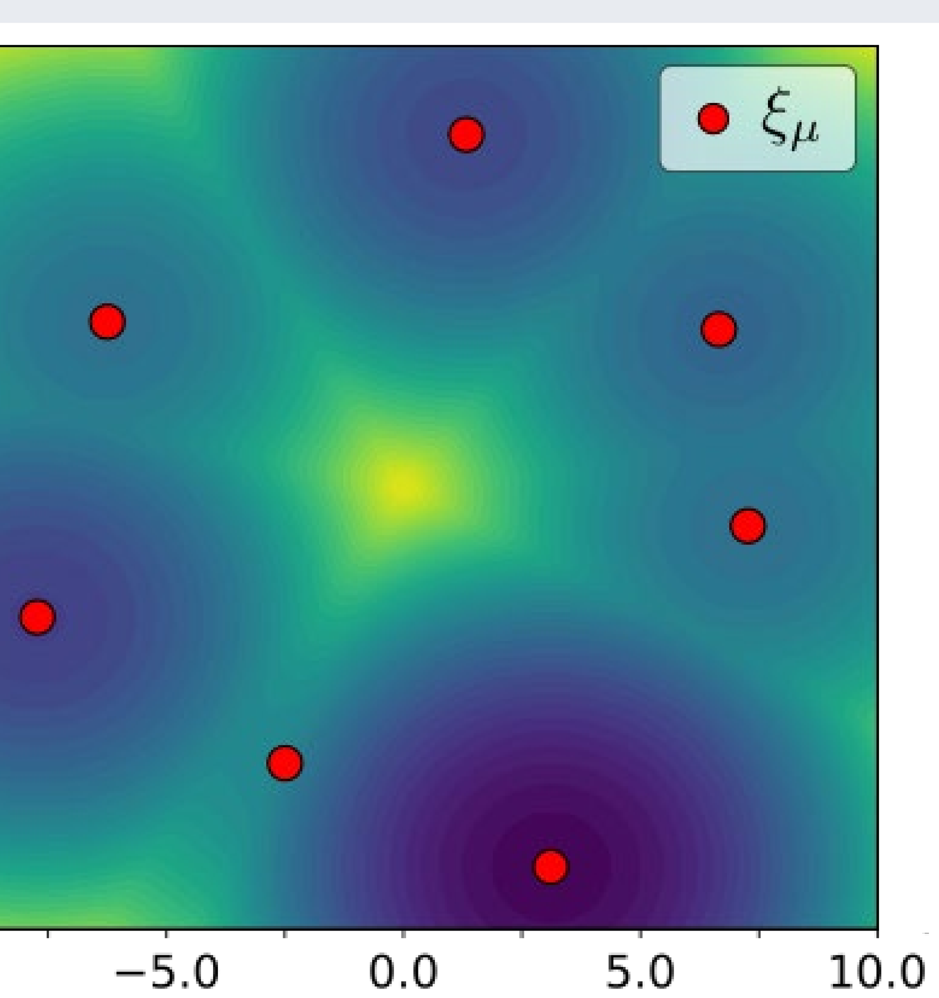


te Update rule

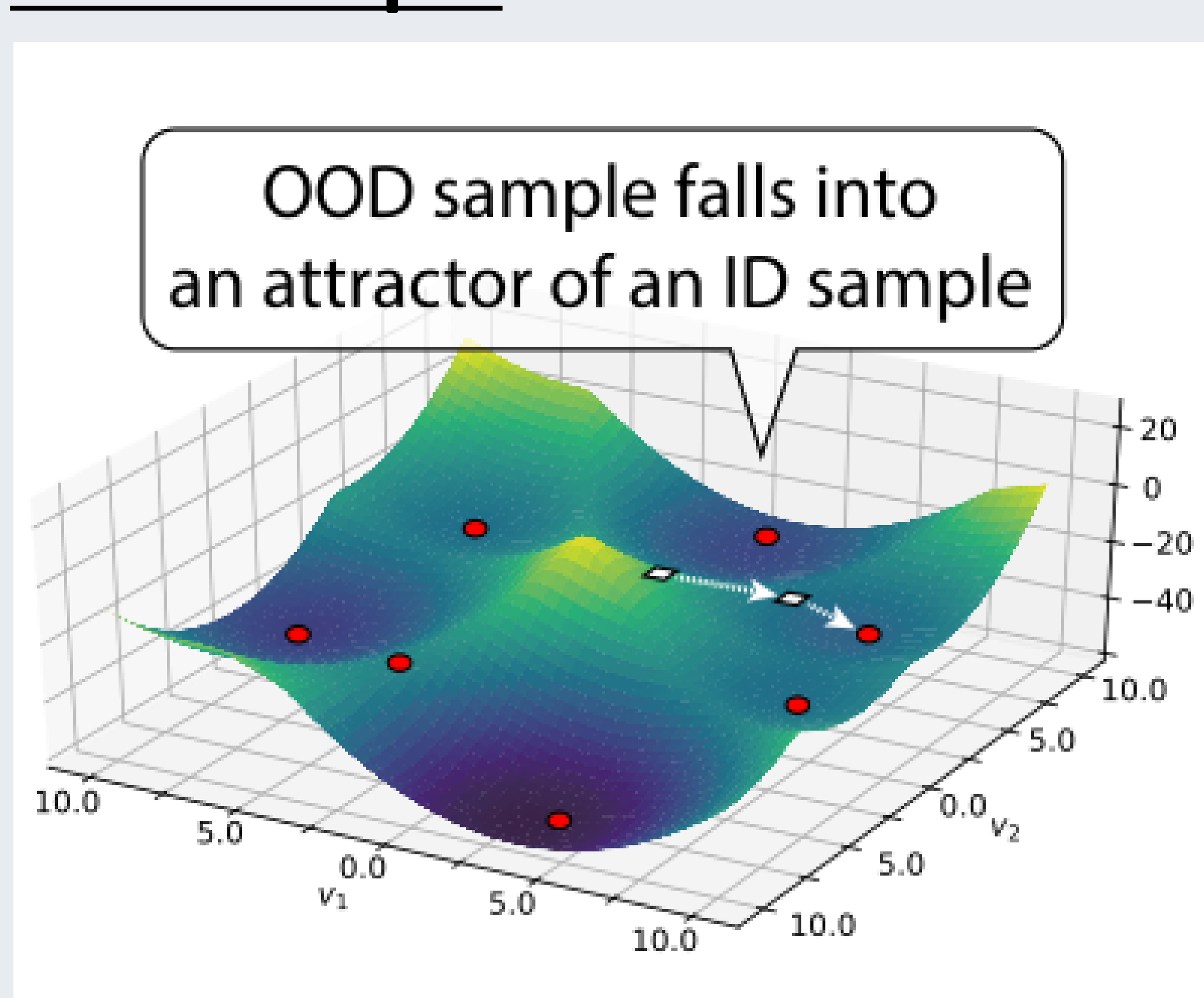
$$v^{k+1} = \Xi^\top \mathrm{softmax}(\beta \Xi v^k)$$

$\in \mathbb{R}^{N_V}$: visible layer neuron $h \in \mathbb{R}^{N_H}$: hidden layer neuron, β: real hyperparameter,
$\Xi \in \mathbb{R}^{N_H \times N_V}$: matrix parameter, ξ_μ: μ-th column vector component of a matrix Ξ

y Map



OOD sample



$$L_H(h) = \max\left(\frac{1}{\beta}\log\left(\frac{1}{\gamma}\sum_{\mu=1} e^{\beta h_\mu}\right), 0\right)$$

Update rule

$$v^{k+1} = \chi\left(G(v^k)\right)\Xi^\top \mathrm{softmax}(\beta \Xi\, v^k)$$

$$G(v) = \frac{1}{\beta}\log\left(\frac{1}{\gamma}\sum_{\mu=1}^{N_H} e^{\beta \xi_\mu v}\right) \qquad \chi(x) = \begin{cases} 1\ (x \geq 0) \\ 1\ (x < 0) \end{cases}$$

$\gamma \in \mathbb{R}_{\geq 0}$: Real hyperparameter

Probability-based learning

W is determined through optimization to satisfy the following probability equa

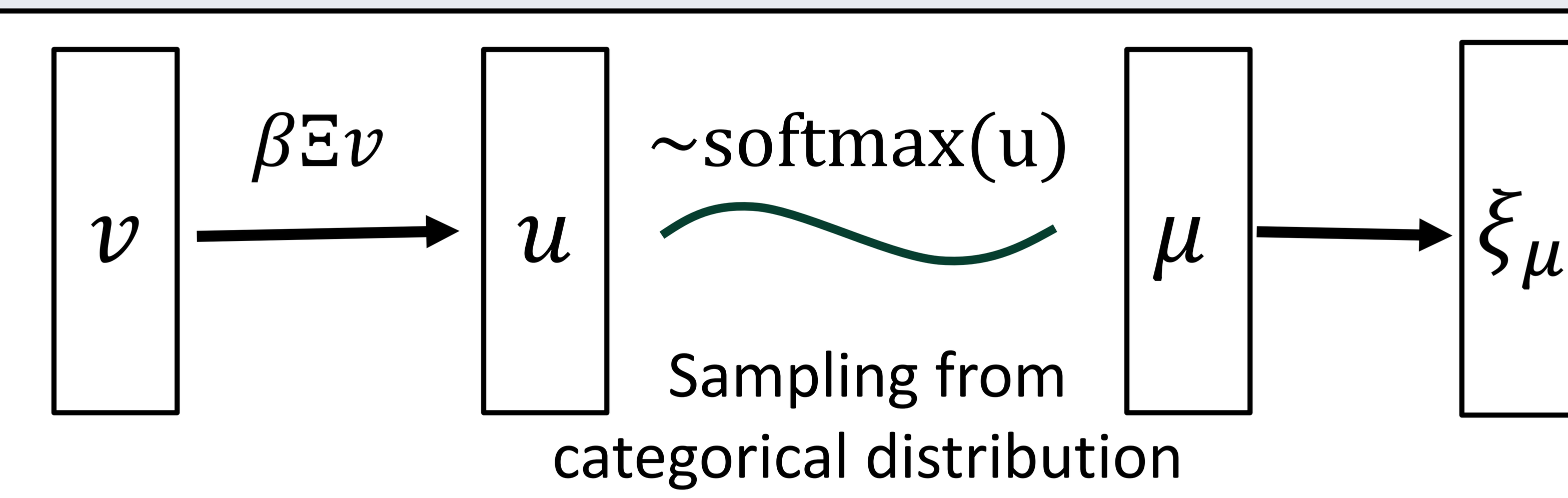
$$p_H(\mu, v) = \frac{1}{Z} e^{\beta \xi_\mu v} \propto f(h)_\mu$$

Each neuron in the hidden laye corresponds to the probability μ-th memory

$$\log(p(v)) - \log\left(\sum_{\mu=1}^{N_H} e^{\beta \xi_\mu v}\right) = \mathrm{const}$$

The Lagrangian of the hidden la corresponds to the probability relative to the memory's paren distribution.

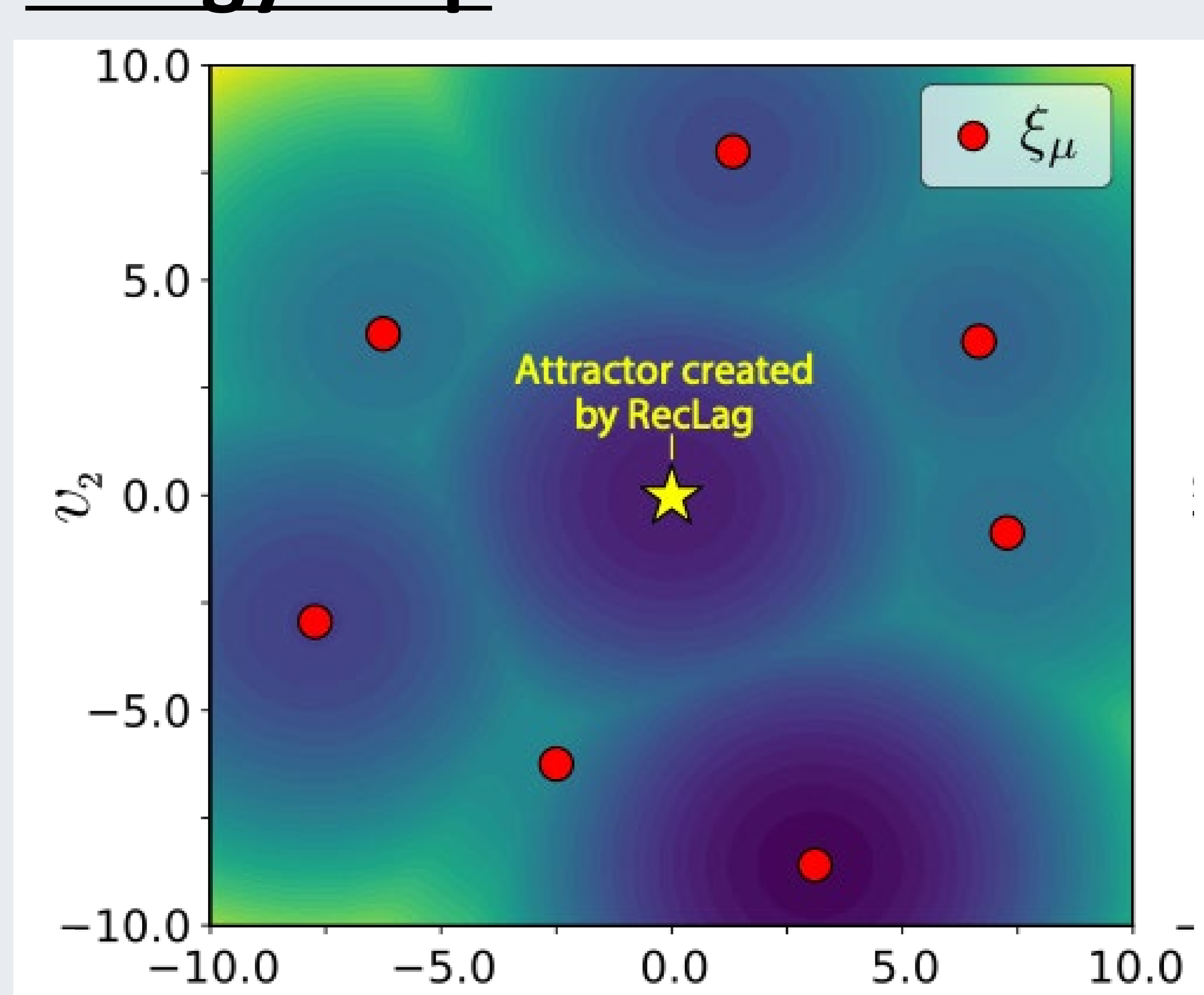
$Z \in \mathbb{R}$: normalized coefficient



$$W^* = \arg\min_W \mathcal{N}(v|\xi_\mu;$$

$\mathcal{N}$: multivariate normal distribution $I_{M\times M}$: M-dimensional identity matrix

Energy Map



OOD sample