Learning with Hyperspherical Uniformity

Weiyang Liu* Rongmei Lin* Zhen Liu* Li Xiong
Bernhard Schölkopf Adrian Weller

AISTATS 2021

Background

- Two types of regularizations for neural networks:

$$\mathcal{L}_{\text{reg}} = \underbrace{\lambda_{\text{I}} \cdot \sum_{i=1}^{n} f(\boldsymbol{w}_i)}_{\text{Individual Regularization}} + \underbrace{\lambda_{\text{R}} \cdot g(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)}_{\text{Relational Regularization}}$$

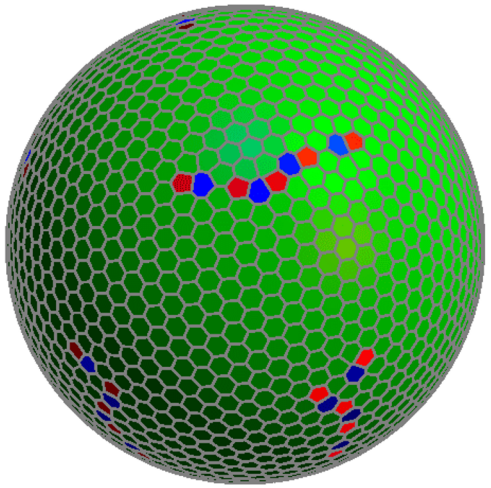
E.g. Weight Decay

E.g. Orthogonality
Hyperspherical Uniformity

- Relational regularization is very important to neural representation and generalization.
- Hyperspherical Uniformality, the angular diversity of neurons on hyperspheres, proves helpful.

Minimum Hyperspherical Energy

- Constructing **repulsion forces** between any pair of weight vectors (in every layer)
- It connects to **Thomson problem** - to find a minimal energy configuration of electrons (with the existence of Coulomb's law) on the surface of an atom.



Liu, Lin, Liu, Liu, Yu, Dai, Song. Learning towards Minimum Hyperspherical Energy, NeurIPS 2018

Minimum Hyperspherical Energy

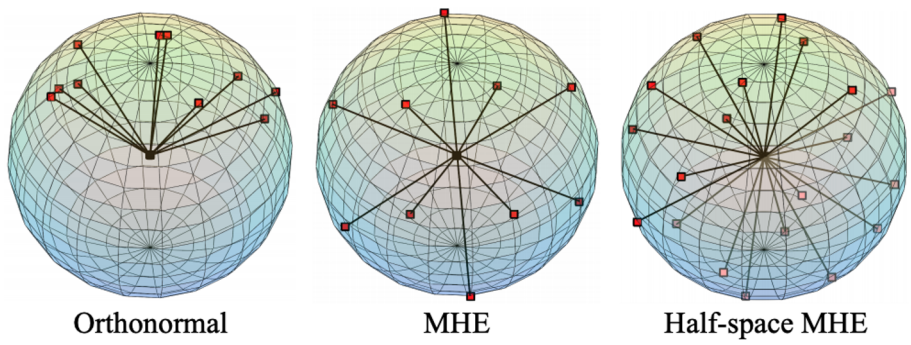
- The objective is formulated as

$$\min_{\{\hat{\boldsymbol{w}}_1,\cdots,\hat{\boldsymbol{w}}_n\in\mathbb{S}^{d-1}\}} \{E_s(\hat{\boldsymbol{W}}_n) := \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K_s(\hat{\boldsymbol{w}}_i,\hat{\boldsymbol{w}}_j)\}$$

Serve as repulsion forces

Riesz s-kernel $K_s(\hat{\boldsymbol{w}}_i,\hat{\boldsymbol{w}}_j) = \begin{cases} \rho(\hat{\boldsymbol{w}}_i,\hat{\boldsymbol{w}}_j)^{-s}, & s>0 \\ \log(\rho(\hat{\boldsymbol{w}}_i,\hat{\boldsymbol{w}}_j)^{-1}), & s=0 \\ -\rho(\hat{\boldsymbol{w}}_i,\hat{\boldsymbol{w}}_j)^{-s}, & s<0 \end{cases}$

Normalized L2 distance or angular distance



Our work aims to answer the following questions

- What is the connection between hyperspherical uniformity and orthogonality?
- Is there any other way to achieve hyperspherical uniformity?
- Is there a unified view on hyperspherical uniformity?

Connection between Hyperspherical Uniformity and Orthogonality

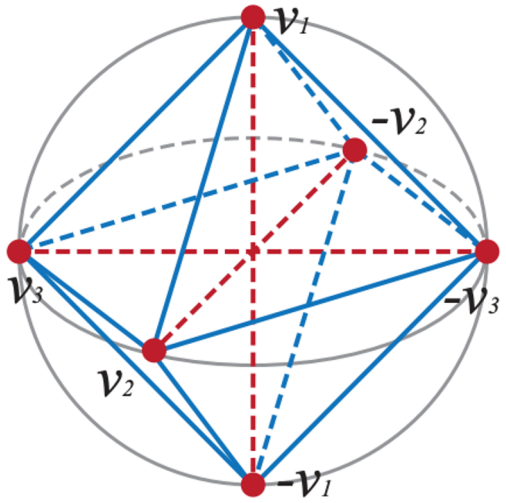
- An Intriguing Example
  - Promoting orthogonality of the following *d+1* vectors:

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{d+1} \in \mathbb{S}^d\}$$

  - Promoting hyperspherical uniformity for the following *2d+2* vectors:

$$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{d+1}, -\boldsymbol{v}_1, \cdots, -\boldsymbol{v}_{d+1} \in \mathbb{S}^d\}$$

  - They are equivalent.



Cross-polytope

Variant: Maximum Hyperspherical Separation

- Inspired by **Tammes problem** where one packs a given number of circles on the surface of a sphere such that the minimum distance between circles is maximized
- Achieving hyperspherical uniformity from a local perspective

$$\max_{\{\hat{\boldsymbol{w}}_1,\cdots,\hat{\boldsymbol{w}}_n \in \mathbb{S}^{d-1}\}} \left\{ \vartheta(\hat{\boldsymbol{W}}_n) := \min_{i \neq j} \rho(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j) \right\}$$

Normalized Euclidean distance or angular distance

Variant: Maximum Hyperspherical Polarization

- MHP maximizes the following s-polarization:

$$\max_{\{\hat{\boldsymbol{w}}_1,\cdots,\hat{\boldsymbol{w}}_n\in\mathbb{S}^{d-1}\}} \left\{P_s(\hat{\boldsymbol{W}}_n) := \min_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \sum_{i=1}^{n} K_s(\boldsymbol{v},\hat{\boldsymbol{w}}_i)\right\}$$

Riesz s-kernel

- It is a max-min problem and amounts to identifying the optimal location of “poles” for the potential function

Variant: Minimum Hyperspherical Covering

- MHC minimizes the following covering radius:

$$\min_{\{\hat{\boldsymbol{w}}_1,\cdots,\hat{\boldsymbol{w}}_n \in \mathbb{S}^{d-1}\}} \left\{ \alpha(\hat{\boldsymbol{W}}_n) := \max_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \min_{1\geq i\geq n} \rho(\boldsymbol{v}, \hat{\boldsymbol{w}}_i) \right\}$$

Normalized L2 distance or angular distance

- The covering radius ***α*** denotes the maximum geodesic distance from a point ***v*** to the nearest point in ***W***.
- It can be viewed as the geodesic radius of the largest hyperspherical cap that contains no points on the hypersphere

Variant: Maximum Gram Determinant

- MGD maximizes the following kernel Gram determinant:

$$\max_{\{\hat{\boldsymbol{w}}_1,\cdots,\hat{\boldsymbol{w}}_n \in \mathbb{S}^{d-1}\}} \log\det\left(\boldsymbol{G} := \left(K(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j)\right)_{i,j=1}^{n}\right)$$

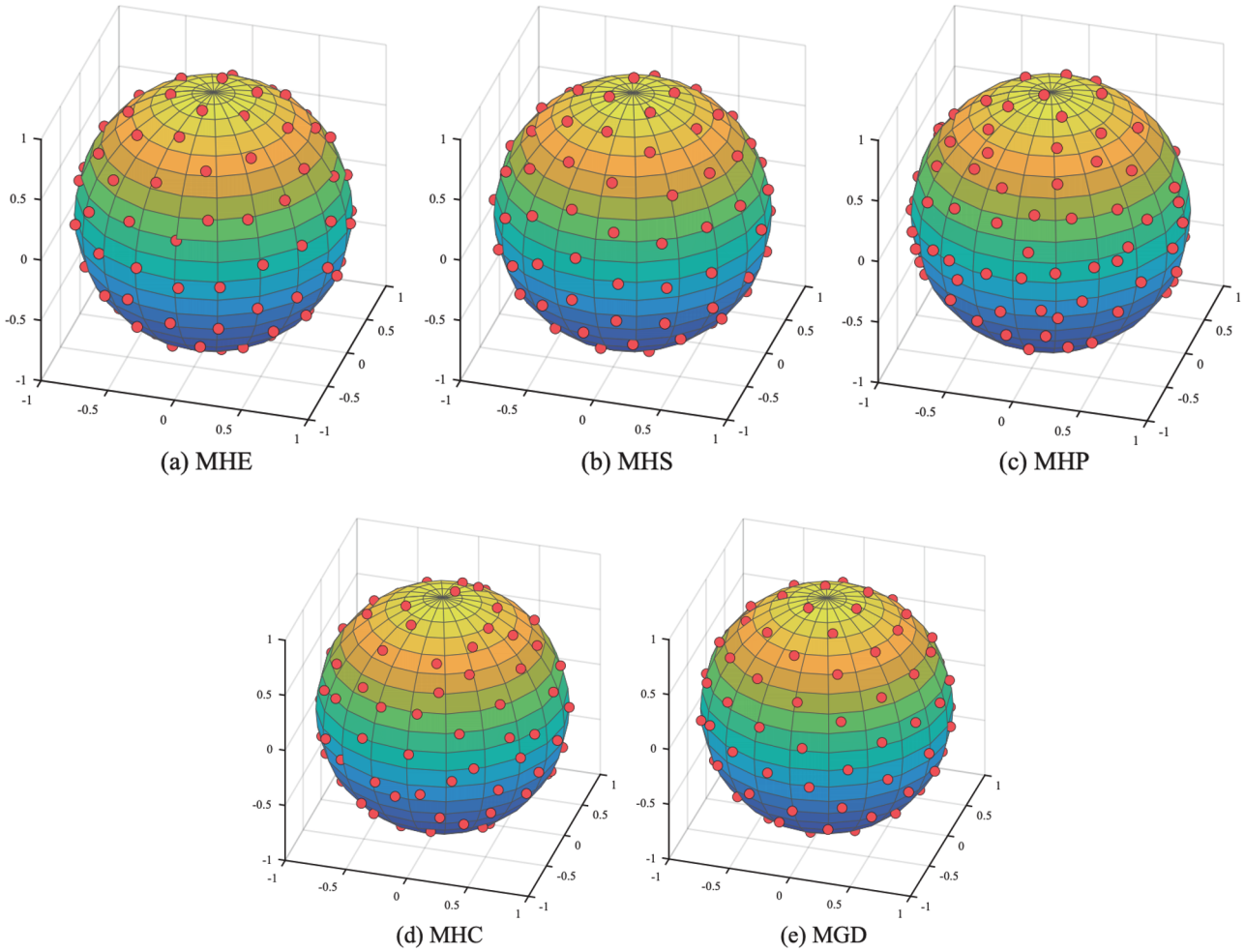
Gaussian Kernel $K(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-\sum_{i=1}^{d} \epsilon_i^2 (u_i - v_i)^2\right)$

- It is inspired by numerical integration and interpolation and also known as extremal systems in numerical integration.
- Geometrically, the kernel Gram determinant is also closely related to n-dimensional volume of the parallelotope formed bt ***W***.

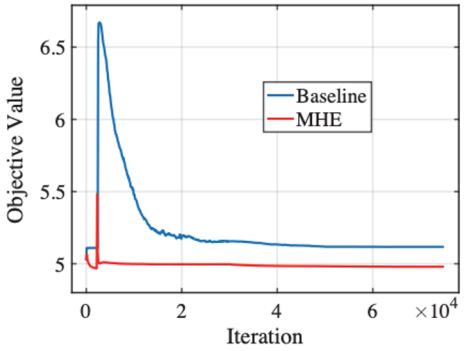
Theoretical Properties (informal)

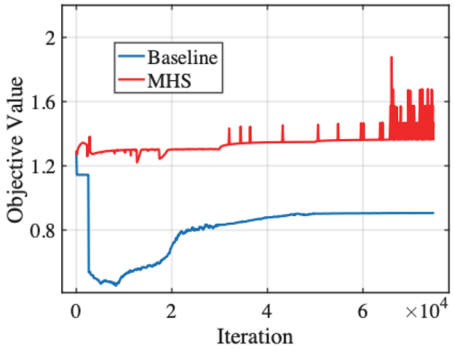
- All these regularizations are asymptotically approaching to uniform spherical measure (uniform distribution on hypersphere).
- All these regularizations are highly connected. For example, MHS is a special case of MHE when s goes to infinity.
- Hyperspherical uniformity yields constrained spectral property.
- **Statistical uniformity testing** on hypersphere can serve as a unified framework to understand the proposed regularizations.
- MHP and MHC are inherently difficult to optimize due to the max-min (min-max) formulation.

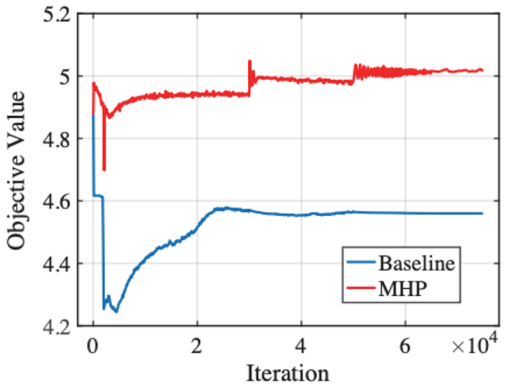
Regularization Effects (3D visualization)

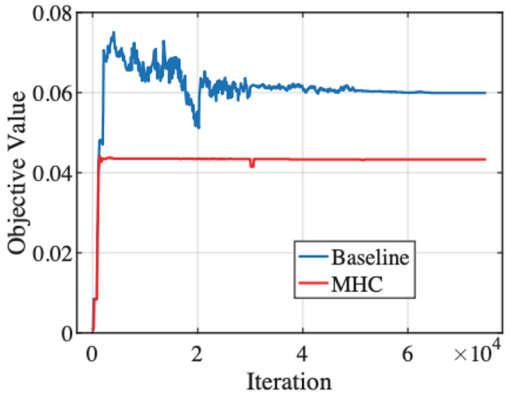


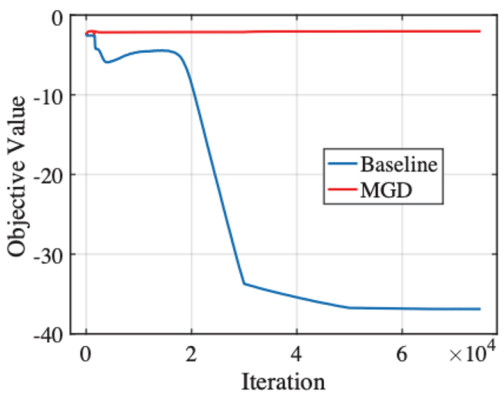
Regularization Effects (Objective Value)











Discriminative Learning

Method	CNN-9	ResNet-18
Baseline	28.13	22.87
Orthogonal	26.94	22.36
MHE	25.94	21.82
MHS	25.43	**20.97**
MHP	25.92	21.24
R-MHP	26.02	22.19
MHC	25.62	21.88
MGD	**25.32**	21.06

CNN on CIFAR-100

Method	Error
Baseline	32.95
Orthogonal	32.65
SRIP	32.53
MHE	32.45
MHS	**32.06**
MHP	32.32
R-MHP	32.71
MHC	32.28
MGD	32.16

CNN on CIFAR-100

Method	Citeseer	Cora	Pubmed
Baseline	70.3	81.3	79.0
Orthogonal	70.4	81.5	78.8
MHE	71.5	82.0	79.0
MHS	71.7	**82.3**	**79.2**
MHP	71.3	81.5	79.0
MHC	71.2	81.6	79.0
MGD	**71.8**	**82.3**	**79.2**

Graph Convolution Networks

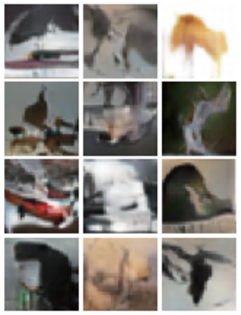
Method	Accuracy
Baseline	87.10
MHE	87.44
MHS	87.60
MHP	87.41
R-MHP	87.10
MHC	87.33
MGD	**87.61**

Point Cloud Networks

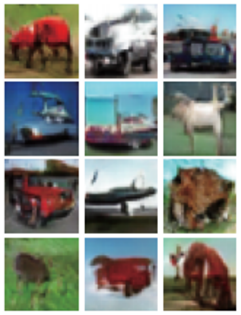
Generative Modeling

- Regularizing GAN on CIFAR-10

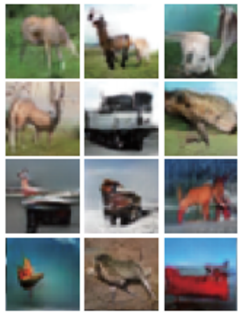
Method	Inception Score
Baseline	7.14
SN [47]	7.40
MHE	7.40
MHS	**7.61**
R-MHP	7.31
MGD	7.49



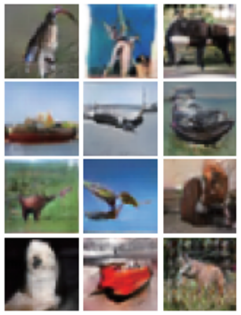
Baseline



MHE



MGD



MHS