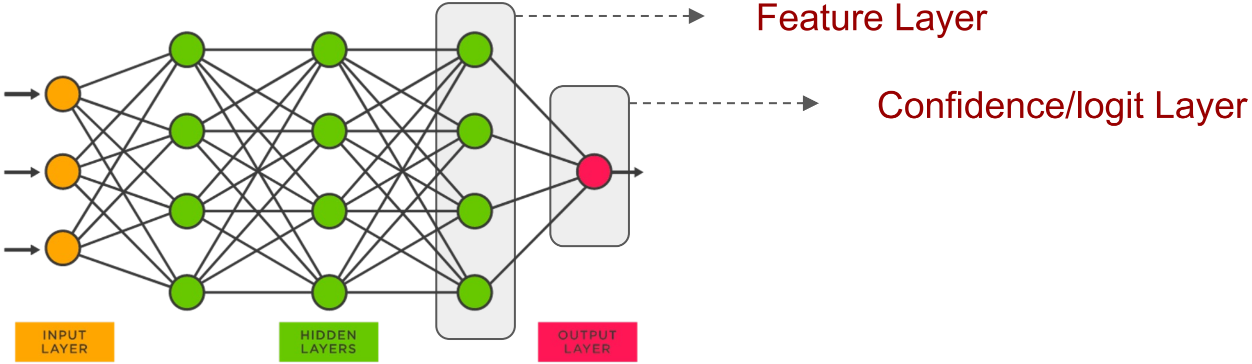
Generalizing and Decoupling Neural Collapse via Hyperspherical Uniformity Gap

Weiyang Liu*, Longhui Yu*, Adrian Weller, Bernhard Schölkopf

ICLR 2023

What is Neural Collapse (NC)?

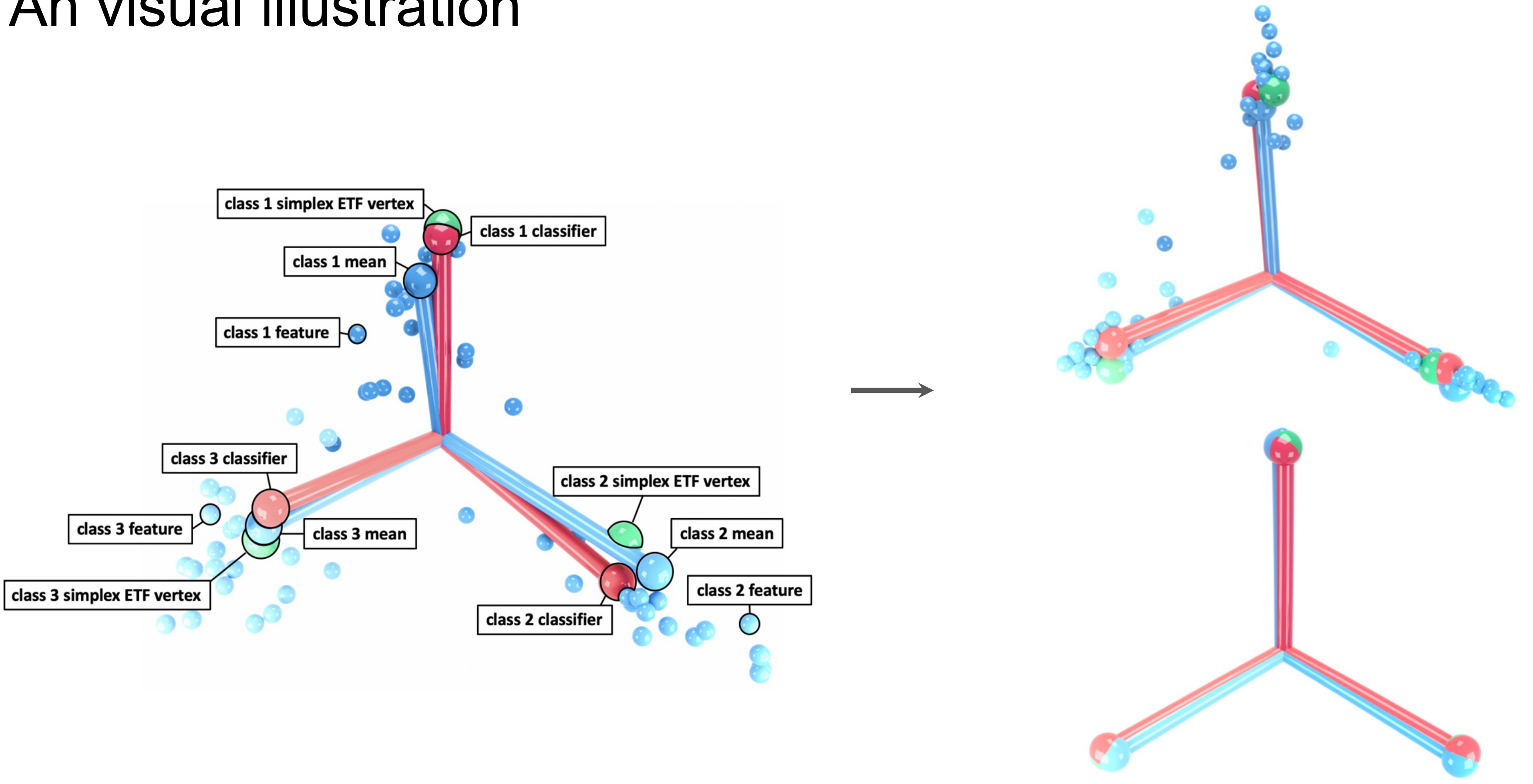
- Modern practice for training neural networks involves a terminal phase of training (TPT), which begins at the epoch where training error first vanishes.
- During TPT, the training error stays effectively zero, while training loss is pushed toward zero.
- TPT exposes a pervasive symmetry and geometric inductive bias, called neural collapse



What is Neural Collapse?

- **Intra-class variability collapse**: Intra-class variability of last-layer features collapses to zero, indicating that all the features of the same class concentrate to their intra-class feature mean.
- **Convergence to simplex ETF**: After being centered at their global mean, the class-means form a simplex equiangular tight frame (ETF) which is a symmetric structure defined by a set of maximally distant and pair-wise equiangular points on a hypersphere.
- **Convergence to self-duality**: The linear classifiers, which live in the dual vector space to that of the class-means, converge to their corresponding class-mean and also form a simplex ETF.
- **Nearest decision rule**: The linear classifiers behave like nearest class-mean classifiers.

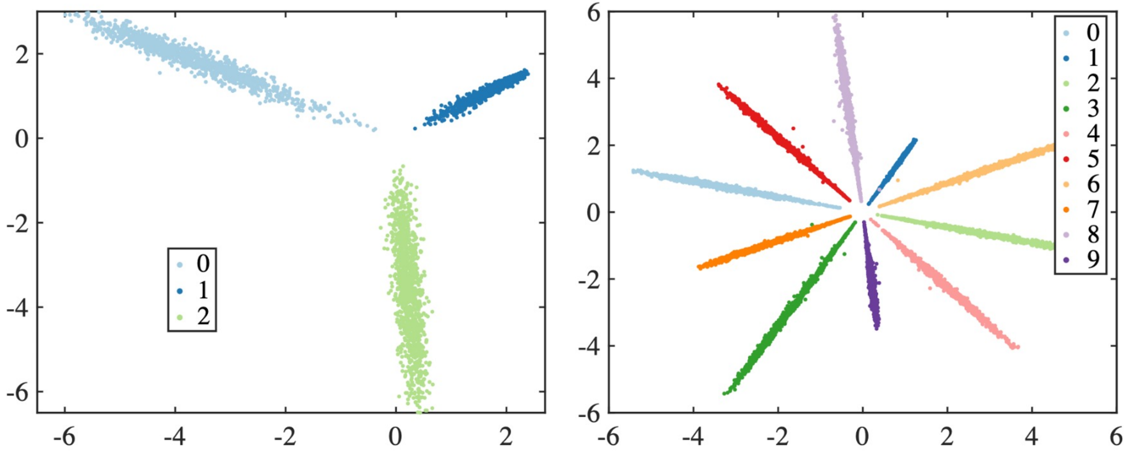
An visual illustration



The Pitfall of Neural Collapse

- Simplex ETF does NOT exist when **the number of classes (C) is larger than the dimension of feature (d)**, but such a scenario is ubiquitous in practice, e.g., contrastive self-supervised learning, extreme classification, face recognition, etc.

Train CNN with cross entropy and set feature dimension as 2 on MNIST



(a) 2D feature with 3 classes

d=2, C=3

(b) 2D feature with 10 classes

d=2, C=10

Generalized Neural Collapse (GNC)

- **Intra-class variability collapse:** $\boldsymbol{\Sigma}_B^{\dagger}\boldsymbol{\Sigma}_W \rightarrow \mathbf{0}$ $\quad$ $\boldsymbol{\Sigma}_W = \text{Ave}_{i,c}(\boldsymbol{x}_{i,c}-\boldsymbol{\mu}_c)(\boldsymbol{x}_{i,c}-\boldsymbol{\mu}_c)^\top$ $\quad$ $\boldsymbol{\Sigma}_B = \text{Ave}_c(\boldsymbol{\mu}_c-\boldsymbol{\mu}_G)(\boldsymbol{\mu}_c-\boldsymbol{\mu}_G)^\top$
- **Convergence to hyperspherical uniformity**: After being centered at their global mean, the class-means are maximally distant on a hypersphere:

$$\sum_{c\neq c'} K(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\mu}}_{c'}) \rightarrow \min_{\hat{\boldsymbol{\mu}}_1,\cdots,\hat{\boldsymbol{\mu}}_C} \sum_{c\neq c'} K(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\mu}}_{c'}), \quad \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\| - \|\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\| \rightarrow 0, \ \forall c \neq c'$$

$$\hat{\boldsymbol{\mu}}_i = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_G\|^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_G)$$

where K is a kernel function and here we consider Riesz s-kernel

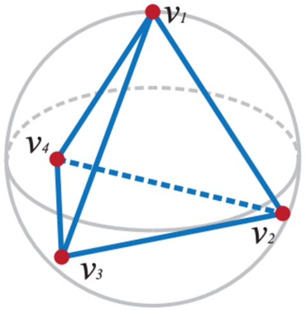
$$K_s(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\mu}}_{c'}) = \text{sign}(s)\cdot\|\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}}_{c'}\|^{-s}$$

- **Convergence to self-duality**: $\|\boldsymbol{w}_c\|^{-1}\boldsymbol{w}_c - \hat{\boldsymbol{\mu}}_c \rightarrow 0$ $\quad$ where *w* denotes the classifier.
- **Nearest decision rule**: $\arg\max_c \langle \boldsymbol{w}_c, \boldsymbol{x}\rangle + b_c \rightarrow \arg\min_c \|\boldsymbol{x} - \boldsymbol{\mu}_c\|$

GNC Provably Covers NC

- Simplex ETF is a global optimum for GNC:

Theorem 1 (Regular Simplex Optimum for GNC) *Let $f:(0,4]\to\mathbb{R}$ be a convex and decreasing function defined at $v=0$ by $\lim_{v\to 0^+} f(v)$. If $2\le C\le d+1$, then we have that the vertices of regular $(C-1)$-simplices inscribed in $\mathbb{S}^{d-1}$ with centers at the origin (equivalent to simplex ETF) minimize the hyperspherical energy $\sum_{c\neq c'} K(\hat{\boldsymbol{\mu}}_c,\hat{\boldsymbol{\mu}}_{c'})$ on the unit hypersphere $\mathbb{S}^{d-1}$ $(d\ge 3)$ with the kernel as $K(\hat{\boldsymbol{\mu}}_c,\hat{\boldsymbol{\mu}}_{c'})=f(\|\hat{\boldsymbol{\mu}}_c-\hat{\boldsymbol{\mu}}_{c'}\|^2)$. If f is strictly convex and strictly decreasing, then these are the only energy minimizing C-point configurations. Thus GNC reduces to NC when $d\ge C-1$.*

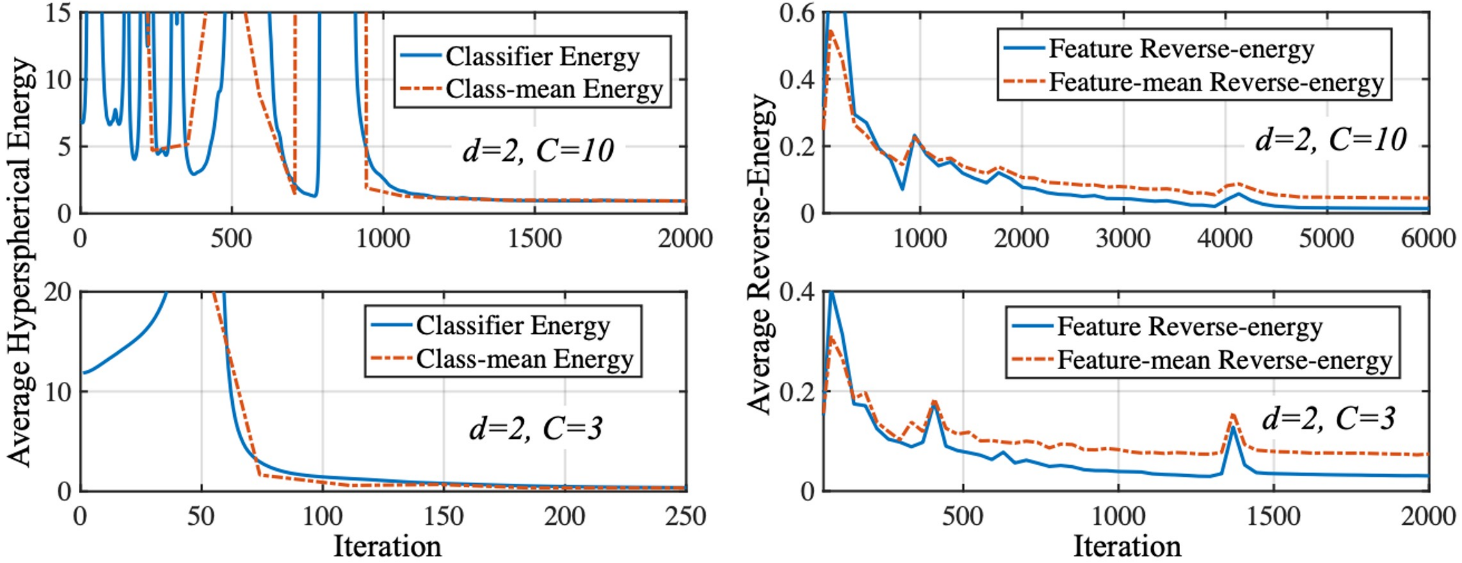


Regular Simplex

Why GNC is Interesting?

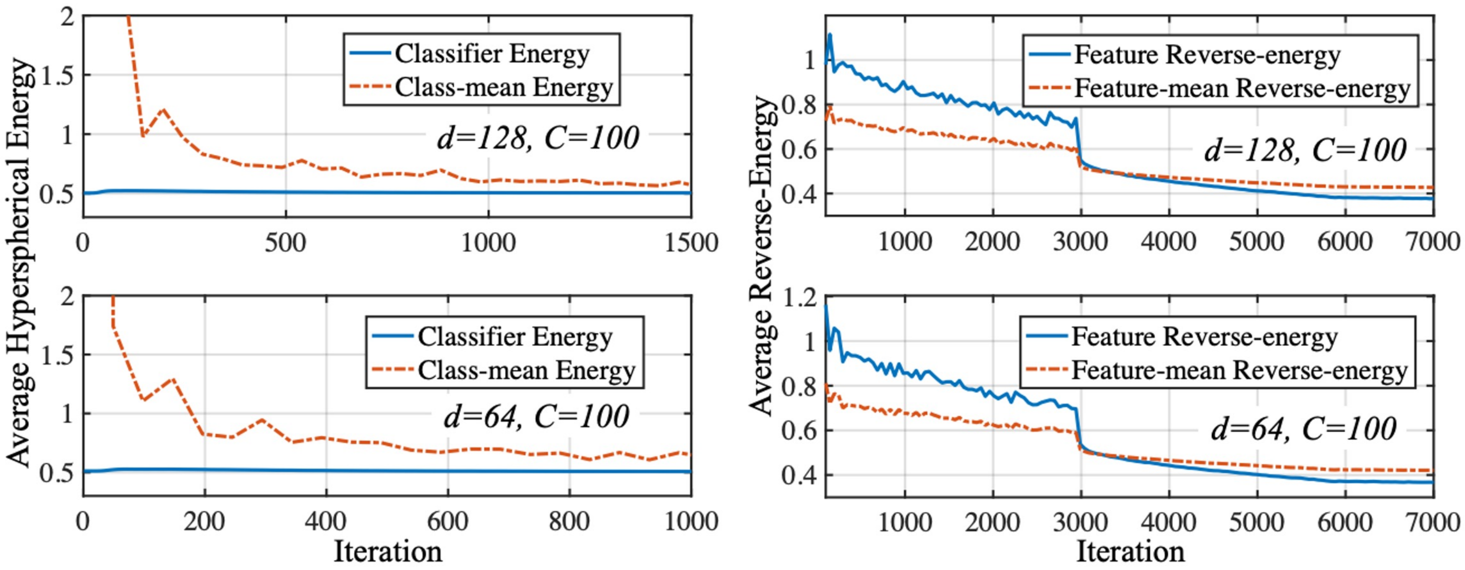
- GNC fully covers the case of NC, while being able to generalize to the case of $d<C$.
- Similar to NC that connects frame theory to deep learning, GNC connects potential theory to deep learning.
- We use a variational characterization of hyperspherical uniformity, which is easily optimizable and gives us natural learning objective (unlike NC).
- We can prove that the widely used cross-entropy loss also converges to GNC.

Empirical Evidence to Validate GNC



(a) Inter-class separability on MNIST (b) Intra-class variability on MNIST

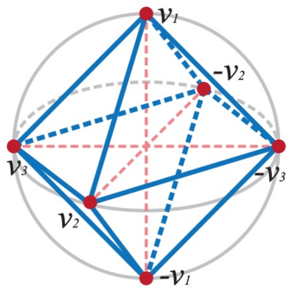
Empirical Evidence to Validate GNC



(c) Inter-class separability on CIFAR-100 (d) Intra-class variability on CIFAR-100

The same empirical phenomenon also happens in ResNet / ViT on ImageNet!

More Theoretical Results on GNC



Cross-Polytope

Theorem 2 (Cross-polytope Optimum for GNC) *If $C = 2d$, then the vertices of the cross-polytope are the minimizer of the hyperspherical energy in GNC(2).*

Theorem 3 (Asymptotic Convergence to Hyperspherical Uniformity) *Consider a sequence of point configurations $\{\hat{\boldsymbol{\mu}}_1^C, \cdots, \hat{\boldsymbol{\mu}}_C^C\}_{C=2}^{\infty}$ that asymptotically minimizes the hyperspherical energy on $\mathbb{S}^{d-1}$ as $C \to \infty$, then $\{\hat{\boldsymbol{\mu}}_1^C, \cdots, \hat{\boldsymbol{\mu}}_C^C\}_{C=2}^{\infty}$ is uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.*

Decoupling GNC: A New Loss Function

- The cross-entropy (CE) loss is arguably the *de facto* choice for classification loss function.
- While we have proved that CE can provably achieve GNC, it also couples two independent criteria: intra-class variability – GNC(1) and inter-class separability – GNC(2).
- GNC shows that these two criteria can be fully decoupled and learned separately, which yields more flexibility.
- With the characterization of uniformity, we identify a quantity called Hyperspherical Uniformity Gap (HUG) that serves as an alternative loss function other than CE

Hyperspherical Uniformity Gap

- General version

$$\max_{\{\hat{\boldsymbol{x}}_i\}_{i=1}^n} \mathcal{L}_{\text{HUG}} := \alpha \cdot \underbrace{\mathcal{HU}\left(\{\hat{\boldsymbol{\mu}}_c\}_{c=1}^C\right)}_{T_b:\text{ Inter-class Hyperspherical Uniformity}} - \beta \cdot \sum_{c=1}^{C} \underbrace{\mathcal{HU}\left(\{\hat{\boldsymbol{x}}_i\}_{i\in A_c}\right)}_{T_w:\text{ Intra-class Hyperspherical Uniformity}}$$

provably minimizing $\mathcal{I}(\widehat{Z};Y)$ = $\mathcal{H}(\widehat{Z})$ - $\mathcal{H}(\widehat{Z}|Y)$

- Proxy-based version (with classifiers)

$$\max_{\{\hat{\boldsymbol{x}}_i\}_{i=1}^n,\{\hat{\boldsymbol{w}}_c\}_{c=1}^C} \mathcal{L}_{\text{P-HUG}} := \alpha \cdot \underbrace{\mathcal{HU}\left(\{\hat{\boldsymbol{w}}_c\}_{c=1}^C\right)}_{\text{Inter-class Hyperspherical Uniformity}} - \beta \cdot \sum_{c=1}^{C} \underbrace{\mathcal{HU}\left(\{\hat{\boldsymbol{x}}_i\}_{i\in A_c}, \hat{\boldsymbol{w}}_c\right)}_{\text{Intra-class Hyperspherical Uniformity}}$$

Variational Characterization of Hyperspherical Uniformity

- For the function *HU*, we consider the following choices:
  - Minimizing the potential energy:

$$\min_{\{\hat{\boldsymbol{v}}_1,\cdots,\hat{\boldsymbol{v}}_n\in\mathbb{S}^{d-1}\}}\left\{E_s(\hat{\boldsymbol{V}}_n):=\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K_s(\hat{\boldsymbol{v}}_i,\hat{\boldsymbol{v}}_j)\right\} \qquad K_s(\hat{\boldsymbol{v}}_i,\hat{\boldsymbol{v}}_j)=\begin{cases}\|\hat{\boldsymbol{v}}_i-\hat{\boldsymbol{v}}_j\|^{-s}, & s>0\\ -\|\hat{\boldsymbol{v}}_i-\hat{\boldsymbol{v}}_j\|^{-s}, & s<0\end{cases}$$

  - Maximizing the separation distance:

$$\max_{\hat{\boldsymbol{V}}}\{\vartheta(\hat{\boldsymbol{V}}_n):=\min_{i\neq j}\|\hat{\boldsymbol{v}}_i-\hat{\boldsymbol{v}}_j\|\}$$

  - Maximum gram determinant:

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

Some Simple Variants from the HUG Framework

- From minimizing the potential energy:

$$\mathcal{L}'_{\text{MHE-HUG}} = \alpha \cdot \sum_{c \neq c'} \|\hat{\boldsymbol{w}}_c - \hat{\boldsymbol{w}}_{c'}\|^{-2} + \beta' \cdot \sum_{c} \sum_{i \in A_c} \|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{w}}_c\|$$

- From maximizing the separation distance:

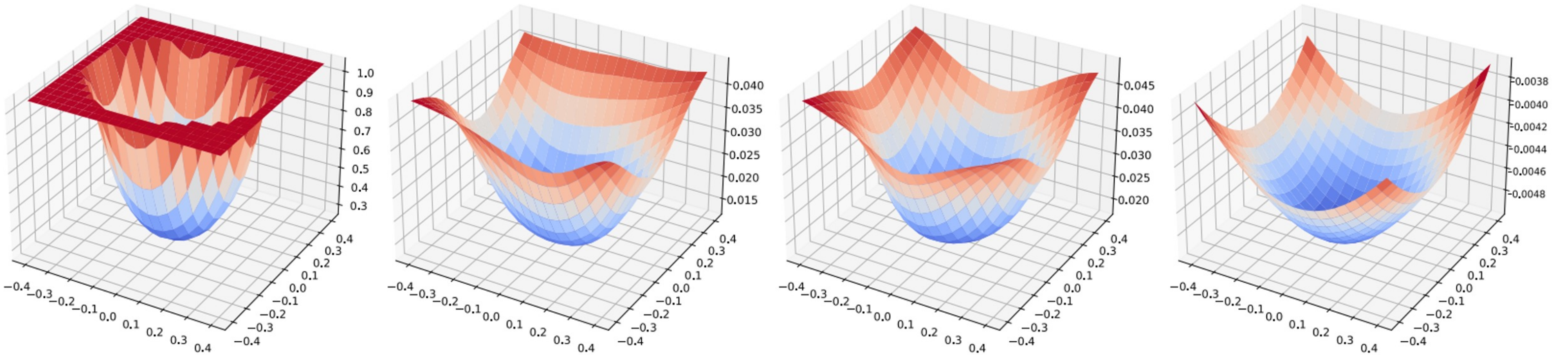
$$\mathcal{L}'_{\text{MHS-HUG}} := \alpha \cdot \min_{c \neq c'} \|\hat{\boldsymbol{w}}_c - \hat{\boldsymbol{w}}_{c'}\| - \beta \cdot \sum_{c} \max_{i \in A_c} \|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{w}}_c\|$$

- From maximizing the gram determinant:

$$\mathcal{L}_{\text{MGD-HUG}} := \alpha \cdot \log\det\left(\boldsymbol{G}(\{\hat{\boldsymbol{w}}_c\}_{c=1}^{C})\right) + \beta' \cdot \sum_{c} \sum_{i \in A_c} \|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{w}}_c\|$$

Loss Landscape Visualization

- More smooth and convex loss landscape



(a) CE Loss (b) HUG: Both losses (c) HUG: Intra-class Variability (d) HUG: Inter-class Separability

Decoupled Loss Function Enables Flexibility

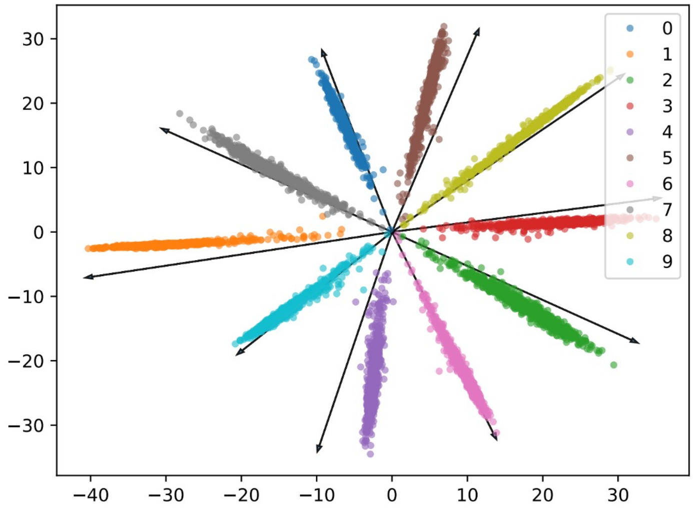
- Learning last-layer classifiers is effortless

Method	CIFAR-10	CIFAR-100
CE Loss	5.45	24.90
Fully learnable	**5.03**	**23.50**
Static (random)	5.19	24.23
Static (optimized)	5.12	24.02
Partially learnable	5.08	23.89

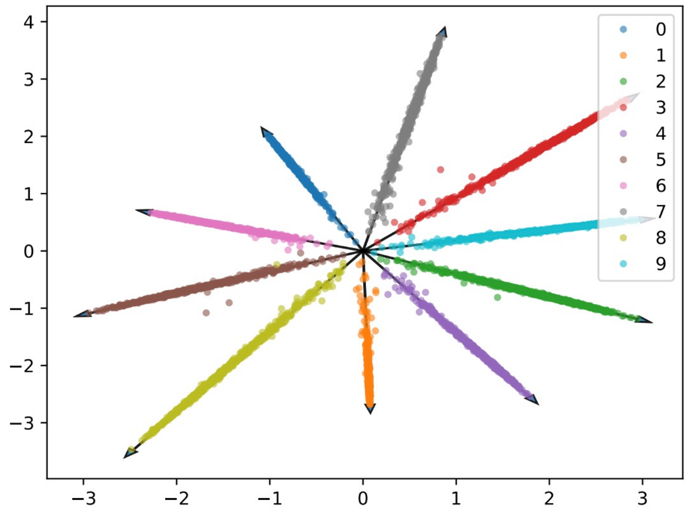
- The performance gain is agnostic to network architectures

Method	ResNet-18	VGG-16	DenseNet-121
CE Loss	5.45 / 24.90	5.28 / 22.99	5.04 / 21.47
HUG	**5.03 / 23.50**	**5.19 / 22.77**	**4.85 / 21.30**

Visualization of learned features



Cross-entropy loss



HUG loss

Experiments

- Better OOD generalization and robustness

	CIFAR-100				CIFAR-10			
IR	0.2	0.1	0.02	0.01	0.2	0.1	0.02	0.01
CE	66.74	62.31	48.79	43.82	90.29	87.85	79.17	74.11
HUG	**67.83**	**63.33**	**50.48**	**45.63**	**90.41**	**88.20**	**79.88**	**75.14**

Long-tail Recognition

	CIFAR-100			CIFAR-10		
Memory size	200	500	2000	200	500	2000
ER + CE	22.14	31.02	43.54	49.07	61.58	76.89
ER + HUG	**23.52**	**31.92**	**43.92**	**53.74**	**62.67**	**77.21**

Continual Learning

Method	Clean	l_∞=2/255	l_∞=4/255	l_∞=8/255
CE Loss	5.45 / 24.90	7.94 / 2.12	0.61 / 0	0 / 0
HUG	**5.03 / 23.50**	**15.24 / 5.26**	**3.45 / 1.24**	**1.76 / 0.44**

Adversarial Robustness