GENERALIZING AND DECOUPLING NEURAL COLLAPSE VIA HYPERSPHERICAL UNIFORMITY GAP

Weiyang Liu$^{1,2,*}$, Longhui Yu$^{3,*}$, Adrian Weller$^{2,4}$, Bernhard Schölkopf$^4$

$^1$Max Planck Institute for Intelligent Systems - Tübingen | $^2$University of Cambridge
$^3$Peking University | $^4$The Alan Turing Institute

ABSTRACT

The neural collapse (NC) phenomenon describes an underlying geometric symmetry for deep neural networks, where both deeply learned features and classifiers converge to a simplex equiangular tight frame. It has been shown that both cross-entropy loss and mean square error can provably lead to NC. We remove NC’s key assumption on the feature dimension and the number of classes, and then present a generalized neural collapse (GNC) hypothesis that effectively subsumes the original NC. Inspired by how NC characterizes the training target of neural networks, we decouple GNC into two objectives: minimal intra-class variability and maximal inter-class separability. We then use hyperspherical uniformity (which characterizes the degree of uniformity on the unit hypersphere) as a unified framework to quantify these two objectives. Finally, we propose a general objective – hyperspherical uniformity gap (HUG), which is defined by the difference between inter-class and intra-class hyperspherical uniformity. HUG not only provably converges to GNC, but also decouples GNC into two separate objectives. Unlike cross-entropy loss that couples intra-class compactness and inter-class separability, HUG enjoys more flexibility and serves as a good alternative loss function. Empirical results show that HUG works well in terms of generalization and robustness.

1 INTRODUCTION

Recent years have witnessed the great success of deep representation learning in a variety of applications ranging from computer vision [37], natural language processing [16] to game playing [55, 64]. Despite such a success, how deep representations can generalize to unseen scenarios and when they might fail remain a black box. Deep representations are typically learned by a multi-layer network with cross-entropy (CE) loss optimized by stochastic gradient descent. In this simple setup, [86] has shown that zero loss can be achieved even with arbitrary label assignment. After continuing to train the neural network past zero loss with CE, [60] discovers an intriguing phenomenon called neural collapse (NC). NC can be summarized as the following characteristics:

• **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 are both linearly separable and maximally distant on a hypersphere. Formally, 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.

The NC phenomenon suggests two general principles for deeply learned features and classifiers: minimal intra-class compactness of features (i.e., features of the same class collapse to a single point), and maximal inter-class separability of classifiers / feature mean (i.e., classifiers of different classes have maximal angular margins). While these two principles are largely independent, popular loss functions such as CE and square error (MSE) completely couple these two principles together. Since there is no trivial way for CE and MSE to decouple these two principles, we identify a novel quantity – hyperspherical uniformity gap (HUG), which not only characterizes intra-class feature compactness and inter-class classifier separability as a whole, but also fully decouples these two
principles. The decoupling enables HUG to separately model intra-class compactness and inter-class separability, making it highly flexible. More importantly, HUG can be directly optimized and used to train neural networks, serving as an alternative loss function in place of CE and MSE for classification. HUG is formulated as the difference between inter-class and intra-class hyperspherical uniformity. Hyperspherical uniformity [48] quantifies the uniformity of a set of vectors on a hypersphere and is used to capture how diverse these vectors are on a hypersphere. Thanks to the flexibility of HUG, we are able to use many different formulations to characterize hyperspherical uniformity, including (but not limited to) minimum hyperspherical energy (MHE) [45], maximum hyperspherical separation (MHS) [48] and maximum gram determinant (MGD) [48]. Different formulations yield different interpretation and optimization difficulty (e.g., HUG with MHE is easy to optimize, HUG with MGD has interesting connection to geometric volume), thus leading to different performance.

Similar to CE loss, HUG also provably leads to NC under the setting of unconstrained features [53]. Going beyond NC, we hypothesize a generalized NC (GNC) with hyperspherical uniformity, which extends the original NC to the scenario where there is no constraint for the number of classes and the feature dimension. NC requires the feature dimension no smaller than the number of classes while GNC no longer requires this. We further prove that HUG also leads to GNC at its objective minimum.

Another motivation behind HUG comes from the classic Fisher discriminant analysis (FDA) [19] where the basic idea is to find a projection matrix $T$ that maximizes between-class variance and minimizes within-class variance. What if we directly optimize the input data (without any projection) rather than optimizing the linear projection in FDA? We make a simple derivation below:

**Projection FDA:** $\max_{T \in \mathbb{R}^{d \times r}} \text{tr} \left( \left( T^\top S_b T \right)^{-1} T^\top S_w T \right)$

**Data FDA:** $\max_{x_1, \ldots, x_n \in \mathbb{R}^{d-1}} \text{tr} \left( S_b - S_w \right)$

where the between-class scatter matrix is $S_b = \sum_{i=1}^C \sum_{j \in A_i} (x_j - \mu)(x_j - \mu)^\top$, the within-class scatter matrix is $S_w = \sum_{i=1}^C n_i (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top$, the total number of classes is $n$, the number of samples in the $i$-th class is $n_i$, $\mu_i = \frac{1}{n_i} \sum_{j \in A_i} x_j$, the within-class mean is $\bar{\mu}$, and $\bar{\mu} = \frac{1}{n} \sum_{i=1}^C \sum_{j \in A_i} x_j$ is the global mean. By considering class-balanced data on the unit hypersphere, optimizing data FDA is equivalent to simultaneously maximizing $\text{tr}(S_b)$ and minimizing $\text{tr}(S_w)$. Maximizing $\text{tr}(S_b)$ encourages inter-class separability and is a necessary condition for hyperspherical uniformity. Minimizing $\text{tr}(S_w)$ encourages intra-class feature collapse, reducing intra-class variability. Therefore, HUG can be viewed a generalized FDA criterion for learning maximally discriminative features.

However, one may ask the following questions: Why is HUG useful if we already have the FDA criterion? Could we simply optimize data FDA? In fact, the FDA criterion has many degenerate solutions. For example, we consider a scenario of 10-class balanced data where all features from the first 5 classes collapse to the north pole on the unit hypersphere and features from the rest 5 classes collapse to the south pole on the unit hypersphere. In this case, $\text{tr}(S_w)$ is already minimized since it achieves the minimum zero. $\text{tr}(S_b)$ also achieves its maximum $n$ at the same time. In contrast, HUG naturally generalizes FDA without having these degenerate solutions and serves as a more reliable criterion for training neural networks. We summarize our contributions below:

- **We decouple the NC phenomenon into two separate learning objectives:** maximal inter-class separability (i.e., maximally distant class feature mean and classifiers on the hypersphere) and minimal intra-class variability (i.e., intra-class features collapse to a single point on the hypersphere).

- **Based on the two principled objectives induced by NC, we hypothesize the generalized NC which generalizes NC by dropping the constraint on the feature dimension and the number of classes.**

- **We identify a general quantity called hyperspherical uniformity gap, which well characterizes both inter-class separability and intra-class variability.** Different from the widely used CE loss, HUG naturally decouples both principles and thus enjoys better modeling flexibility.

- **Under the HUG framework, we consider three different choices for characterizing hyperspherical uniformity:** minimum hyperspherical energy, maximum hyperspherical separation and maximum Gram determinant. HUG provides a unified framework for using different characterizations of hyperspherical uniformity to design new loss functions.

---

1. We first obtain the upper bound $n$ of $\text{tr}(S_b)$ from $\text{tr}(S_b) = \sum_{i=1}^C n_i \|\mu_i - \bar{\mu}\| \leq \sum_{i=1}^C n_i \|\mu_i\| \leq n$. Because a set of vectors $\{\mu_i\}_{i=1}^n$ achieving hyperspherical uniformity has $\mathbb{E}_{\mu_1, \ldots, \mu_n} \|\bar{\mu}\| \to 0$ (as $n$ grows larger) [25]. Then we have that $\text{tr}(S_b)$ attains $n$. Therefore, vectors achieving hyperspherical uniformity are one of its maximizers. $\text{tr}(S_w)$ can simultaneously attain its minimum if intra-class features collapse to a single point.
2 On Generalizing and Decoupling Neural Collapse

NC describes an intriguing phenomenon for the distribution of last-layer features and classifiers in overly-trained neural networks, where both features and classifiers converge to ETF. However, ETF can only exist when the feature dimension \( d \) and the number of classes \( C \) satisfy \( d \geq C - 1 \). This is not always true for deep neural networks. For example, neural networks for face recognition are usually trained by classifying large number of classes (e.g., more than 85K classes in [23]), and the feature dimension (e.g., 512 in SphereFace [43]) is usually much smaller than the number of classes. In general, when the number of classes is already large, it is prohibitive to use a larger feature dimension. Thus a question arises: what will happen in this case if a neural network is fully trained?

Motivated by this question, we conduct a simple experiment to simulate the case of \( d \geq C - 1 \) and the case of \( d < C - 1 \). Specifically, we train a convolutional neural network (CNN) on MNIST with feature dimension 2. For the case of \( d \geq C - 1 \), we use only 3 classes (digit 0,1,2) as the training set. For the case of \( d < C - 1 \), we use all 10 classes as the training set. We visualize the learned features of both cases in Figure 1. The results verify the case of \( d \geq C - 1 \) indeed approaches to NC, and ETF does not exist in the case of \( d < C - 1 \). Interestingly, one can observe that learned features in both cases approach to the configuration of equally spaced frames on the hypersphere. To accommodate the case of \( d < C - 1 \), we extend NC to the generalized NC by hypothesizing that last-layer inter-class features and classifiers converge to equally spaced points on the hypersphere, which can be characterized by hyperspherical uniformity.

---

**Generalized Neural Collapse (GNC)**

We define the feature global mean as \( \mu_G = \text{Ave}_{i,c} x_{i,c} \) where \( x_{i,c} \in \mathbb{R}^d \) is the last-layer feature of the \( i \)-th sample in the \( c \)-th class, the feature class-mean as \( \mu_c = \text{Ave}_{i,c} x_{i,c} \) for different classes \( c \in \{1, \ldots, C\} \), the feature within-class covariance as \( \Sigma_W = \text{Ave}_{i,c}(x_{i,c} - \mu_c)(x_{i,c} - \mu_c)^\top \) and the feature between-class covariance as \( \Sigma_B = \text{Ave}_c(\mu_c - \mu_G)(\mu_c - \mu_G)^\top \). GNC states that

- **(1) Intra-class variability collapse**: Intra-class variability of last-layer features collapse to zero, indicating that all the features of the same class converge to their intra-class feature mean. Formally, GNC has that \( \Sigma_B^\dagger \Sigma_W \to 0 \) where \( \dagger \) denotes the Moore-Penrose pseudoinverse.

- **(2) Convergence to hyperspherical uniformity**: After being centered at their global mean, the class-means are both linearly separable and maximally distant on a hypersphere. Formally, the class-means converge to equally spaced points on a hypersphere, i.e.,

  \[
  \sum_{c \neq c'} K(\hat{\mu}_c, \hat{\mu}_{c'}) \to \min_{\mu_1 \ldots \mu_C} \sum_{c \neq c'} K(\hat{\mu}_c, \hat{\mu}_{c'}) \quad \|\mu_c - \mu_{c'}\| \to 0, \forall c \neq c'
  \]

  where \( \hat{\mu}_c = \|\mu_c - \mu_G\|^{-1}(\mu_c - \mu_G) \) and \( K(\cdot, \cdot) \) is a kernel function that models pairwise interaction. Typically, we consider Riesz \( s \)-kernel \( K_s(\mu_c, \mu_{c'}) = \text{sign}(s) \|\mu_c - \mu_{c'}\|^{-s} \) or logarithmic kernel \( K_{\log}(\mu_c, \mu_{c'}) = \log \|\mu_c - \mu_{c'}\|^{-1} \). For example, the Riesz \( s \)-kernel with \( s = d - 2 \) is a variational characterization of hyperspherical uniformity (e.g., hyperspherical energy [45]) using Newtonian potentials. In the case of \( d = 3, s = 1 \), the Riesz kernel is called Coulomb potential and the problem of finding minimal coulomb energy is called Thomson problem [70].

- **(3) 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-means, leading to hyperspherical uniformity. Formally, GNC has that \( \|w_c\|^{-1} w_c - \mu_c \to 0 \) where \( w_c \in \mathbb{R}^d \) is the \( c \)-th classifier.

- **(4) Nearest decision rule**: The learned linear classifiers behave like the nearest class-mean classifiers. Formally, GNC has that \( \arg \max_c \langle w_c, x \rangle + b_c \to \arg \min_c \|x - \mu_c\| \).

In contrast to NC, GNC further considers the case of \( d < C - 1 \) and hypothesizes that both feature class-means and classifiers converge to hyperspherically uniform point configuration that minimizes some form of pairwise potentials. Similar to how NC connects tight frame theory [74] to deep learning, our GNC hypothesis connects potential theory [3] to deep learning, which may shed new light on understanding it. We show in Theorem 1 that GNC reduces to NC in the case of \( d \geq C - 1 \).
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 $2 \leq C \leq d + 1$, then we have that the vertices of regular $(C - 1)$-simplices inscribed in $S^{d-1}$ with centers at the origin (equivalent to simplex ETF) minimize the hyperspherical energy $E = \sum_{c \neq c'} K(\mu_c, \mu_{c'})$ on the unit hypersphere $S^{d-1} (d \geq 3)$ with the kernel as $K(\mu_c, \mu_{c'}) = f(\|\mu_c - \mu_{c'}\|^2)$. If $f$ is strictly convex and strictly decreasing, then these are the only energy minimizing $C$-point configurations. Thus GNC reduces to NC when $d \geq C - 1$.

We note that Theorem 2 guarantees the simplex ETF as the minimizer of a general family of hyperspherical energies (as long as $f$ is convex and decreasing). This suggests that there are many possible kernel functions $K(\cdot, \cdot)$ in GNC that can effectively generalize NC. The case of $d < C - 1$ is where GNC really gets interesting but complicated. Other than the regular simplex case, we also highlight a special uniformity case of $2d = C$. In this case, we can prove in Theorem 2 that GNC(2) converges to the vertices of a cross-polytope as hyperspherical energy gets minimized. As the number of classes gets infinitely large, we show in Theorem 3 that GNC(2) leads to a point configuration that is uniformly distributed on $S^{d-1}$. Additionally, we show a simple yet interesting result in Proposition 1 that the last-layer classifiers are already initialized to be uniformly distributed on the hypersphere in practice.

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).

The cross-polytope optimum for GNC(2) is in fact quite intuitive, because it corresponds to the Cartesian coordinate system (up to a rotation). For example, the vertices of the unit cross-polytope in $\mathbb{R}^3$ are $(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)$. These 6 vectors minimize the hyperspherical energy on $S^2$. We illustrate both the regular simplex and cross-polytope cases in Figure 2. For the other cases of $d < C - 1$, there exists generally no simple and universal point structure that minimizes the hyperspherical energy, as heavily studied in [12, 27, 38, 63]. For the point configurations that asymptotically minimize the hyperspherical energy as $C$ grows larger, Theorem 3 can guarantee that these configurations asymptotically converge to a uniform distribution on the hypersphere.

Theorem 3 (Asymptotic Convergence to Hyperspherical Uniformity) Consider a sequence of point configurations $\{\mu_1^C, \ldots, \mu_C^C\}_{C=1}^{\infty}$ that asymptotically minimizes the hyperspherical energy on $S^{d-1}$ as $C \to \infty$, then $\{\mu_1^C, \ldots, \mu_C^C\}_{C=1}^{\infty}$ is uniformly distributed on the hypersphere $S^{d-1}$.

Proposition 1 (Minimum Energy Initialization) With zero-mean Gaussian initialization (e.g., [22, 28]), the $C$ last-layer classifiers of neural networks are initialized as a uniform distribution on the hypersphere. The expected initial energy is $E(C) = (d-1) \int_{S^{d-1}} \int_{S^{d-1}} \|\mu_c - \mu_{c'}\|^{-2} \sigma_{d-1}(\mu_c) \sigma_{d-1}(\mu_{c'})$.

With Proposition 1, one can expect that the hyperspherical energy of the last-layer classifiers will first increase and then decrease to a lower value than the initial energy. To validate the effectiveness of our GNC hypothesis, we conduct a few experiments to show how both class feature means and classifiers converge to hyperspherical uniformity (i.e., minimizing the hyperspherical energy), and how intra-class feature variability collapses to almost zero. We start with an intuitive understanding about GNC from Figure 1. The results are directly produced by the learned features without any visualization tool (such as t-SNE [73]), so the feature distribution can reflect the underlying one learned by neural networks. We observe that GNC is attained in both $d < C - 1$ and $d \geq C - 1$, while NC is violated in $d < C - 1$ since the learned feature class-means can no longer form a simplex ETF. To see whether the same conclusion holds for higher feature dimensions, we also train two CNNs on CIFAR-100 with feature dimension as 64 and 128, respectively. The results are given in Figure 3.

Figure 3 shows that GNC captures well the underlying convergence of the neural network training. Figure 3(a,c) shows that the hyperspherical energy of feature class-means and classifiers converge to a small value, verifying the correctness of GNC(2) and GNC(3) which indicate both feature class-means and classifiers converge to hyperspherical uniformity. More interestingly, in the MNIST experiment, we can compute the exact minimal energy on $S^1$; 2 in the case of $d = 2, C = 3$ (1/3 for average energy) and $\approx 82.5$ in the case of $d = 2, C = 10$ ($\approx 0.917$ for average energy). The final average energy in Figure 3(a) matches our theoretical minimum well. From Figure 3(c), we observe that the classifier energy stays close to its minimum at the very beginning, which matches our Proposition 1.
that vectors initialized with zero-mean Gaussian are uniformly distributed over the hypersphere (this phenomenon becomes more obvious in higher dimensions). To evaluate the intra-class feature variability, we consider a hyperspherical reverse-energy $E_r = \sum_{i \neq j, i \in A_c} \|\hat{x}_i - \hat{x}_j\|$ where $\hat{x}_i = \frac{x_i}{\|x_i\|}$ and $A_c$ denotes the sample index set of the $c$-th class. The smaller this reverse-energy gets, the less intra-class variability it implies. Figure 3(b, d) shows that the intra-class feature variability approaches to zero, as GNC(1) suggests. Details and more empirical results on GNC are in given Appendix A.

Now we discuss how to decouple the GNC hypothesis and how such a decoupling can enable us to design new objectives to train neural networks. GNC(1) and GNC(2) suggest to minimize intra-class feature variability and maximize inter-class feature separability, respectively. GNC(3) and GNC(4) are natural consequences if GNC(1) and GNC(2) hold. It has long been discovered in [42, 68, 82] that last-layer classifiers serve as proxies to represent the corresponding class of features, and they are also an approximation to the feature class-means. GNC(3) indicates the classifiers converge to hyperspherical uniformity, which, together with GNC(1), implies GNC(4).

Until now, it has been clear that GNC really boils down to two decoupled objectives: maximize inter-class separability and minimize intra-class variability, which again echoes the goal of FDA. The problem reduces to how to effectively characterize these two objectives while being decoupled for flexibility (unlike CE or MSE). In the next section, we propose to address this problem by characterizing both objectives with a unified quantity - hyperspherical uniformity.

3 Hyperspherical Uniformity Gap

3.1 General Framework

As GNC(2) suggests, the inter-class separability is well captured by hyperspherical uniformity of feature class-means, so it is natural to directly use it as a learning target. On the other hand, GNC(1) does not suggest any easy-to-use quantity to characterize intra-class variability. We note that minimizing intra-class variability is actually equivalent to encouraging features of the same class to concentrate on a single point, which is the opposite of hyperspherical uniformity. Therefore, we can unify both intra-class variability and inter-class separability with a single characterization of hyperspherical uniformity. We propose to maximize the hyperspherical uniformity gap:

$$\max_{\{ x_i \}_{i=1}^C} \mathcal{L}_{\text{HUG}} := \alpha \cdot \frac{\mathcal{H}(\{ \hat{x}_c \}_{c=1}^C)}{T_S}: \text{Inter-class Hyperspherical Uniformity} - \beta \sum_{c=1}^C \frac{\mathcal{H}(\{ \hat{x}_i \}_{i \in A_c})}{T_w}: \text{Intra-class Hyperspherical Uniformity}$$

(2)

where $\alpha, \beta$ are hyperparameters, $\hat{x}_c = \frac{x_c}{\|x_c\|}$ is the feature class-mean projected on the unit hypersphere, $x_c = \sum_{i \in A_c} x_i$ is the feature class-mean, $x_i$ is the last-layer feature of the $i$-th sample and $A_c$ denotes the sample index set of the $c$-th class. $\mathcal{H}(\{ v_{ij} \}_{j=1}^m)$ denotes some measure of hyperspherical uniformity for vectors $\{ v_{1} \cdots v_{m} \}$. Eq. 2 is the general objective for HUG. Without loss of generality, we assume that the larger it gets, the stronger hyperspherical uniformity we have. We mostly focus on supervised learning with parametric class proxies\footnote{Parametric class proxies are a set of parameters used to represent a group of samples in the same class. Therefore, these proxies store the information about a class. Last-layer classifiers are a typical example.} where the CE loss is widely used as a de facto choice, although HUG can be used in much broader settings as discussed later. In the HUG framework, there is no longer a clear notion of classifiers (unlike the CE loss), but we still can utilize class proxies (i.e., a generalized concept of classifiers) to facilitate the optimization.

We observe that Eq. 2 directly optimizes the feature-class means for inter-class separability, but they are intractable to compute during training (we need to compute them in every iteration). Therefore it is nontrivial to optimize the original HUG for training neural networks. A naive solution is to
approximate feature class-mean with a few mini-batches such that the gradients of $T_b$ can be still back-propagated to the last-layer features. However, it may take many mini-batches in order to obtain a sufficiently accurate class-mean, and the approximation gets much more difficult with large number of classes. To address this, we employ parametric class proxies to act as representatives of intra-class features and optimize them instead of feature class-means. We thus modify the HUG objective as

$$\max_{\{\hat{w}_j\}_{j=1}^c} \mathcal{L}_{\text{HUG}} := \alpha \cdot \mathcal{H}(\{\hat{w}_j\}_{j=1}^c) - \beta \cdot \sum_{c=1}^C \mathcal{H}(\{\hat{x}_i\}_{i \in A_c}, \hat{w}_c)$$

where $\hat{w}_c \in S^{d-1}$ is the parametric proxy for the $c$-th class. The intra-class hyperspherical uniformity term connects the class proxies with features by minimizing their joint hyperspherical uniformity, guiding features to move towards their corresponding class proxy. When training a neural network, the objective function in Eq. 3 will optimize network weights and proxies together. There are alternative ways to design the HUG loss from Eq. 2 for different learning scenarios, as discussed in Appendix C.

**Learnable proxies.** We can view the class proxy $\mu_c$ as learnable parameters and update them with stochastic gradients, similarly to the parameters of neural networks. In fact, learnable proxies play a role similar to the last-layer classifiers in the CE loss, improving the optimization by aggregating intra-class features. The major difference between learnable proxies and moving-averaged proxies is the way we update them. As GNC(3) implies, class proxies in HUG can also be used as classifiers.

**Static proxies.** Eq. 3 is decoupled into maximal inter-class separability and minimal intra-class variability. These two objects are independent and do not affect each other. We can thus optimize them independently. This suggests a even simpler way to assign class proxies – initializing class proxies with prespecified points that have attained hyperspherical uniformity, and fixing them in the training. There are two simple ways to obtain these class proxies: (1) minimizing their hyperspherical energy beforehand; (2) using zero-mean Gaussian to initialize the class proxies (Proposition 1). After initialization, class proxies will stay fixed and the features are optimized towards their class proxies.

**Partially learnable proxies.** After the class proxies are initialized using the static way above, we can increase its flexibility by learning an orthogonal matrix for the class proxies to find a suitable orientation for them. Specifically, we can learn this orthogonal matrix using methods in [47].

### 3.2 Variational Characterization of Hyperspherical Uniformity

While there exist many ways to measure hyperspherical uniformity, we seek variational characterization due to simplicity. As examples, we consider minimum hyperspherical energy [45] that is inspired by Thomson problem [66, 70] and minimizes the potential energy, maximum hyperspherical separation [48] that is inspired by Tammes problem [69] and maximizes the smallest pairwise distance, and maximum gram determinant [48] that is defined by the volume of the formed parallelotope.

**Minimum hyperspherical energy.** MHE seeks to find an equilibrium state with minimum potential energy that distributes $n$ electrons on a unit hypersphere as evenly as possible. Hyperspherical uniformity is characterized by minimizing the potential energy for $n$ vectors $V_n = \{v_1, \cdots, v_n \in \mathbb{R}^d\}$:

$$\min_{\{\hat{v}_1, \cdots, \hat{v}_n \in S^{d-1}\}} \left\{ \sum_{i=1}^n \sum_{j=1, j \neq i}^n K_s(\hat{v}_i, \hat{v}_j) \right\}, \quad K_s(\hat{v}_i, \hat{v}_j) = \begin{cases} \|\hat{v}_i - \hat{v}_j\|^{-s}, & s > 0 \\ -\|\hat{v}_i - \hat{v}_j\|^{-s}, & s < 0 \end{cases}, \quad (4)$$

where $\hat{v}_i := \frac{v_i}{\|v_i\|}$ is the $i$-th vector projected onto the unit hypersphere. With $\mathcal{H}(V) = -E_s(V)$, we apply MHE to HUG and formulate the new objective as follows ($s_b = 2, s_w = -1$):

$$\min_{\{\hat{x}_j\}_{j=1}^c, \{\hat{w}_j\}_{j=1}^c} \mathcal{L}_{\text{MHE-HUG}} := \alpha \cdot \mathcal{E}_{\text{MHE}}(\{\hat{w}_j\}_{j=1}^c) - \beta \cdot \sum_{c=1}^C \mathcal{E}_{\text{MHE}}(\{\hat{x}_j\}_{j \in A_c}, \hat{w}_c)$$

which can be relaxed to a upper bound such that we can instead minimize a simple upper bound of $\mathcal{L}_{\text{MHE-HUG}}$:

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

which is much more efficient to compute in practice and thus can serve as a relaxed HUG objective. Moreover, $\mathcal{L}_{\text{MHE-HUG}}$ and $\mathcal{L}'_{\text{MHE-HUG}}$ share the same minimizer. Detailed derivation is in Appendix H.

**Maximum hyperspherical separation.** MHS uses a maximum geodesic separation principle by maximizing the separation distance $\vartheta(V_n)$ (i.e., the smallest pairwise distance in $V_n = \{v_1, \cdots, v_n \in \mathbb{R}^d\}$).
We show in Proposition 4 that CE inherently optimizes two independent criterion: intra-class variability and inter-class separability. We define $\mathcal{H}(\cdot)$ as the intra-class variability term. With MGD, $\epsilon$ which is a max-min optimization with a simple nearest neighbor problem inside. We note that Proposition 4 generally holds with a wide range of $\epsilon$ and $\alpha$. With normalized classifiers and features, we can see how the CE loss implicitly maximizes the HUG objective in a coupled way. GNC indicates that the CE loss eventually approaches to the maximizer of HUG, we now look into compact metric space. We have that Proposition 2 (MHS is a Limiting Case of MHE) Let $n \in \mathbb{N}, n \geq 2$ be fixed and $(S^{d-1}, L_2)$ be a compact metric space. We have that $\lim_{s \to \infty} \frac{1}{n} \min_{\mathcal{V}_n \subset S^{d-1}} E_s(\mathcal{V}_n) = c(s, d)$ where $c(s, d)$ is a constant involving $s, d$. Theorem 4 (Order of Minimum Hyperspherical Energy) If $d - 1 > s > 0$ or $0 > s > -2$ and $d \in \mathbb{N}$, we have that $\lim_{n \to \infty} \{n^{-2} \cdot \min_{\mathcal{V}_n} E_s(\mathcal{V}_n)\} = c(s, d)$ where $c(s, d)$ is a constant involving $s, d$. The result above shows that the leading term of the minimum energy grows of order $\mathcal{O}(n^2)$ as $n \to \infty$. Theorem 4 generally holds with a wide range of $s$ for the Riesz kernel in hyperspherical energy. Moreover, the following result shows that MHS is in fact a limiting case of MHE as $s \to \infty$. Proposition 3 The HUG objectives in both Eq. 5 and Eq. 6 converge to simplex ETF when $2 \leq C \leq d + 1$, converge to cross-polytope when $C = 2d$ and asymptotically converge to GNC as $C \to \infty$. Proposition 3 shows that HUG not only decouples GNC but also provably converges to GNC. Since GNC indicates that the CE loss eventually approaches to the maximizer of HUG, we now look into how the CE loss implicitly maximizes the HUG objective in a coupled way. Proposition 4 The CE loss is $\mathcal{L}_{CE} = \sum_{i=1}^{n} \log(1 + \sum_{j \neq i}^{C} \exp((w_j, x_i) - (w_{y_i}, x_i)))$ where $n$ is the number of samples, $x_i$ is the $i$-th sample with label $y_i$, and $w_i$ is the last-layer linear classifier for the $i$-th class. Bias is omitted for simplicity. $\mathcal{L}_{CE}$ is bounded by $(p = C - 1)$

$$\sum_{i=1}^{n} \sum_{j \neq i}^{C} (w_j, x_i) - \rho \sum_{i=1}^{n} (w_{y_i}, x_i) \leq \mathcal{L}_{CE} \leq \log(1 + \sum_{i=1}^{n} \sum_{j \neq i}^{C} \exp((w_j, x_i)) + \rho \sum_{i=1}^{n} \exp(-(w_{y_i}, x_i)))$$

We show in Proposition 4 that CE inherently optimizes two independent criterion: intra-class variability (IV) and inter-class separability (IS). With normalized classifiers and features, we can see that $Q_1$ and $Q_3$ have similar minimum where $x_i = w_{y_i}$ and $w_1, \forall i$ attain hyperspherical uniformity.
We show that CE is lower bounded by the gap of inter-class and intra-class hyperspherical uniformity:
\[
\mathcal{L}_{CE} \geq \sum_{i=1}^{n} \log \sum_{c=1}^{C} \exp\left(\rho_2 \sum_{j=1}^{n} l_{ic} \langle x_i, x_j \rangle \right) - \rho_3 \sum_{i=1}^{n} \sum_{j \in A_{y_i}} l_{ic} \|x_i - x_j\|^2 - \rho_1 \sum_{i=1}^{n} \sum_{j \in A_{y_i}} \langle x_i, x_j \rangle \tag{11}
\]
where \(\rho_1, \rho_2, \rho_3\) are constants and \(l_{ic}\) is the softmax confidence of \(x_i\) for the \(c\)-th class (Appendix L). This result [4] implies that minimizing CE effectively minimizes HUG. [50] proves that the minimizer of the normalized CE loss converges to hyperspherical uniformity. We rewrite their results below:

**Theorem 5 (CE Asymptotically Converges to HUG’s Maximizer)** Considering unconstrained features of \(C\) classes (each class has the same number of samples), with features and classifiers normalized on some hypersphere, we have, that for the minimizer of the CE loss, classifiers converge weakly to the uniform measure on \(S^{d-1}\) as \(C \rightarrow \infty\) and features collapse to their corresponding classifiers. The minimizer of CE also asymptotically converges to the maximizer of HUG.

Theorem 5 shows that the minimizer of the CE loss with unconstrained features [53] asymptotically converges to the maximizer of HUG (i.e., GNC). Till now, we show that HUG shares the same optimum with CE (with hyperspherical normalization), while being more flexible for decoupling inter-class feature separability and intra-class feature variability. Therefore, we argue that HUG can be an excellent alternative for the widely used CE loss in classification problems.

**HUG maximizes mutual information.** We can view HUG as a way to maximize mutual information \(I(X; Y) = \mathcal{H}(X) - \mathcal{H}(X|Y)\), where \(X\) denotes the feature space and \(Y\) is the label space. Maximizing \(\mathcal{H}(X)\) implies that the feature should be uniform over the space. Minimizing \(\mathcal{H}(X|Y)\) means that the feature from the same class should be concentrated. This is nicely connected to HUG.

**The role of feature and class proxy norm.** Both NC and GNC do not take the norm of feature and class proxy into consideration. HUG also assume both feature and class proxy norm are projected onto some hypersphere. Although dropping these norms usually improves generalizability [8, 9, 14, 43, 76], training neural networks with standard CE loss still yields different class proxy norms and feature norms. We hypothesize that this is due to the underlying difference among training data distribution of different classes. One empirical evidence to support this is that average feature norm of different classes is consistent across training under different random seeds (e.g., average feature norm for digit 1 on MNIST stays the smallest in different run). [36, 46, 52] empirically show that feature norm corresponds to the quality of the sample, which can also viewed as a proxy to sample uncertainty. [56] theoretically shows that the norm of neuron weights (e.g., classifier) matters for its Rademacher complexity. As a trivial solution to minimize the CE loss, increasing the classifier norm (if the feature is correctly classified) can easily decrease the CE loss to zero for this sample, which is mostly caused by the softmax function. Taking both feature and class proxy norm into account greatly complicates the analysis (e.g., it results in weighted hyperspherical energy where the potentials between vectors are weighted) and seem to yield little benefit for now. We defer this issue to future investigation.

**HUG as a general framework for designing loss functions.** HUG can be viewed as an inherently decoupled way of designing new loss functions. As long as we design a measure of hyperspherical uniformity, then HUG enables us to effortlessly turn it into a loss function for neural networks.

4 **EXPERIMENTS AND RESULTS**

Our experiments aims to demonstrate the empirical effectiveness of HUG, so we focus on the fair comparison to the popular CE loss under the same setting. Experimental details are in Appendix N.

4.1 **EXPLORATORY EXPERIMENTS AND ABLATION STUDY**

**Different HUG variants.** We compare different HUG variants and the CE loss on CIFAR-10 and CIFAR-100 with ResNet-18 [29]. Specifically, we use Eq. 6, Eq. 6 and Eq. 10 for MHE-HUG, MHS-HUG and MGD-HUG, respectively. The results are given in Table 1. We can observe that all HUG variants outperform the CE loss. Among all, MHE-HUG achieves the best testing accuracy with considerable improvement over the CE loss. We note that all HUG variants are used without the CE loss. The performance gain of HUG are actually quite significant, since the CE loss is currently a default choice for classification problems and serves as a very strong baseline.
Different methods to update proxies. We also evaluate how different proxy update methods will affect the classification performance. We use the same setting as Table 1. For all the proxy update methods, we apply them to MHE-HUG (Eq. 6) under the same setting. The results are given Table 2. We can observe that all the propose proxy update methods work reasonably well. More interestingly, static proxies work surprisingly well and outperform the CE loss even when all the class proxies are randomly initialized and then fixed throughout the training. The reason the static proxies work for MHE-HUG is due to Proposition 1. This result is significant since we no longer have to train class proxies in HUG (unlike CE). When trained with large number of classes, it is GPU-memory costly for learning class proxies, which is also known as one of the bottlenecks for face recognition [1]. HUG could be a promising solution to this problem.

Loss landscape and convergence. We perturb neuron weights (refer to [40]) to visualize the loss landscape of HUG and CE in Figure 4. We use MHE in HUG here. The results show that HUG yields much flatter local minima than the CE loss in general, implying that HUG has potentially stronger generalization [34, 57]. We show more visualizations and convergence dynamics in Appendix O.

Learning with different architectures. We evaluate HUG with different network architectures such as VGG-16 [65], ResNet-18 [29] and DenseNet-121 [31]. Results in Table 3 (Left number: CIFAR-10, right number: CIFAR-100) show that HUG is agnostic to different network architectures and outperforms the CE loss in every case. Although HUG works well on its own, any other methods that improve CE can also work with HUG.

4.2 Generalization and Robustness under Different Learning Scenarios

Long-tailed recognition. We consider the task of long-tailed recognition, where the data from different classes are imbalanced. The settings generally follow [6], and the dataset gets more imbalanced if the imbalance ratio (IR) gets smaller. The potential of HUG in imbalanced classification is evident, as the inter-class separability in the HUG is explicitly modeled and can be easily controlled. Experimental results in Table 4 show that HUG can consistently outperform the CE loss in the challenging long-tailed setting under different imbalanced ratio.

Continual learning. We demonstrate the potential of HUG in the class-continual learning setting, where the training data is not sampled i.i.d. but comes in class by class. Since training data is highly biased, hyperspherical uniformity among class proxies is crucial. Due to the decoupled nature of HUG, we can easily increase the importance of inter-class separability, unlike CE. We use a simple continual learning method – ER [62] where the CE loss with memory is used. We replace it with HUG. Table 5 shows HUG consistently improves ER under different memory size.

Adversarial robustness. We further test HUG’s adversarial robustness. In our experiments, we consider the classical white-box PGD attack [51] on ResNet-18. The PGD attack iteration is set as 100 and the attack strength level is set as 2/255, 4/255, 8/255 in $l_{\infty}$ norm. All networks are naturally training with either HUG or CE loss. Results in Table 6 demonstrate that HUG yields consistently stronger adversarial robustness than the CE loss.

NLP tasks. As an exploration, we evaluate HUG on some simple NLP classification tasks. Our experiments follow the same settings as [32] and finetune the BERT model [15] in these tasks. Table 7 shows that HUG yields better generalizability than CE, demonstrating its potential for NLP.
5 RELATED WORK AND CONCLUDING REMARKS

We start by generalizing and decoupling the NC phenomenon, obtaining two basic principles for loss functions. Based on these principles, we identify a quantity hyperspherical uniformity gap, which not only decouples NC but also provides a general framework for designing loss functions. We demonstrate a few simple HUG variants that outperform the CE loss in terms of generalization and adversarial robustness. Different from existing work in hyperspherical uniformity \cite{41, 45, 48} and generic diversity (decorrelation) \cite{2, 7, 11, 54, 77, 83}, HUG works as a new learning target (used without CE) rather than acting as a regularizer for the CE loss (used together with CE). Following the spirit of \cite{32}, we demonstrate the effectiveness and potential of HUG as a valid substitute for CE.

**Relevant theoretical results.** \cite{87} has discussed NC under the case of $d < C - 1$, and shown that the global solution in this case yields the best rank-$d$ approximation of the simplex ETF. Along with \cite{87}, GNC gives a more profound characterization of the convergence of class-means. We show a special case of $d = 2, C = 4$. It is easy to see that hyperspherical uniformity in this case forms four vectors with adjacency ones being perpendicular. This is also the case captured by the best rank-2 approximation (i.e., a 2-dimensional hyperplane with simplex ETF projected onto it). Figure 5 gives a geometric interpretation for the connection between \cite{87} and GNC. \cite{3} provides an in-depth introduction and comprehensive theoretical analysis for the energy minimization problem, which significantly benefits this work.

**Connection to contrastive learning.** The goal of contrastive learning \cite{8, 10, 24, 30, 72, 78, 85} is to learn discriminative features through instance-wise discrimination and contrast. Despite the lack of class labels, \cite{78} discovers that contrastive learning performs sample-wise alignment and sample-wise uniformity, sharing a similar high-level spirit to intra-class variability and inter-class separability. \cite{35} adapts contrastive learning to the supervised settings where labeled samples are available, which also shares conceptual similarity to our framework and settings.

**Related work on (deep) metric learning.** Metric learning also adopts similar idea where similar samples are pulled together and dissimilar ones are pushed away. HUG has intrinsic connections to a number of loss functions in metric learning \cite{4, 17, 21, 24, 58, 59, 61, 67, 68, 79, 79–81, 84}.

6 BROADER IMPACT AND FUTURE WORK

Our work reveals the underlying principle – hyperspherical uniformity gap, for classification loss function, especially in the context of deep learning. We provide a simple yet effective framework for designing decoupled classification loss functions. Rather than previous objective functions that are coupled and treated as a black-box, our loss function has clear physical interpretation and is fully decoupled for different functionalities. These characteristics may help neural networks to identify intrinsic structures hidden in data and true causes for classifying images. HUG may have broader applications in interpretable machine learning and fairness / bias problems.

Our work is by no means perfect, and there are many aspects that require future investigation. For example, the implicit data mining in CE \cite{49} is missing in the current HUG design, current HUG losses are more sensitive to hyperparameters than CE (the flexibility of decoupling also comes at a price), current HUG losses could be more unstable to train (more difficult to converge) than CE, and it requires more large-scale experiments to fully validate the superiority of current HUG losses. We hope that our work can serve as a good starting point to rethink classification losses in deep learning.

ACKNOWLEDGEMENT

The authors would like to sincerely thank the anonymous reviewers for all the detailed and valuable suggestions that have significantly improved the paper. This work is supported by the German Federal Ministry of Education and Research (BMBF); Tübingen AI Center, FKZ: 01IS18039A, 01IS18039B; and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. AW acknowledges support from a Turing AI Fellowship under EPSRC grant EP/V025279/1, and the Leverhulme Trust via CFI.
REFERENCES


Appendix

Table of Contents

A Empirical Results on Generalized Neural Collapse 16
   A.1 Detailed Metric Definition ............................ 16
   A.2 Empirical Results of GNC on ImageNet ................. 16

B 2D MNIST Feature Visualization 17

C Other Variants in the HUG Framework 20
   C.1 Proxy-free HUG ........................................ 20
   C.2 Coupled HUG ........................................... 20
   C.3 HUG without Hyperspherical Normalization .......... 21

D Proof of Theorem 1 22

E Proof of Theorem 2 23

F Proof of Theorem 3 24

G Proof of Proposition 1 25

H Derivation of HUG Surrogate for MHE and MHS 26

I Proof of Proposition 2 27

J Proof of Proposition 3 28

K Proof of Proposition 4 29

L Derivation of CE’s Lower Bound 30

M Proof of Theorem 5 31

N Experimental Details 32

O Additional Experimental Results 33
A Empirical Results on Generalized Neural Collapse

A.1 Detailed Metric Definition

We consider four metrics: average classifier energy (ACE), average class-mean energy (ACME), average feature reverse-energy (AFRE) and average feature-mean reverse-energy (AFMRE) in the paper. Their definitions are given below:

\[
E_{\text{ACE}} = \frac{1}{C(C-1)} \sum_{i \neq j} \| \hat{w}_i - \hat{w}_j \|^{-2} \tag{12}
\]

\[
E_{\text{ACME}} = \frac{1}{C(C-1)} \sum_{i \neq j} \| \hat{\mu}_i - \hat{\mu}_j \|^{-2} \tag{13}
\]

\[
E_{\text{AFRE}} = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{|A_c|} \cdot \left( |A_c| - 1 \right) \sum_{i \neq j \in A_c} \| \hat{x}_i - \hat{x}_j \| \tag{14}
\]

\[
E_{\text{AFMRE}} = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{|A_c|} \sum_{i \in A_c} \| \hat{x}_i - \hat{\mu}_c \| \tag{15}
\]

where \(|A_c|\) denotes the cardinality of the set \(A_c\), \(\hat{\mu}_c\) is the normalized feature mean of the \(c\)-th class and \(\hat{w}_c\) denotes the normalized class proxy of the \(c\)-th class.

A.2 Empirical Results of GNC on ImageNet

We find that the GNC hypothesis remains valid and informative even under the scenario of large number of classes (we use the 1000-class ImageNet-2012 dataset [13] here). Experimental results with ResNet-18 [29] (feature dimension as 512) are given in Figure 6. Experimental results with ResNet-50 [29] (feature dimension as 2048) are given in Figure 7.

![Figure 6: Training dynamics of hyperspherical energy (which captures inter-class separability) and hyperspherical reverse-energy (which captures intra-class variability). ImageNet-2012 [13] with ResNet-18 [29] (\(d = 512, C = 1000\)).](image1)

![Figure 7: Training dynamics of hyperspherical energy (which captures inter-class separability) and hyperspherical reverse-energy (which captures intra-class variability). ImageNet-2012 [13] with ResNet-50 [29] (\(d = 2048, C = 1000\)).](image2)
B 2D MNIST Feature Visualization

We also visualize the 2D MNIST feature in Figure 8, Figure 9 and Figure 10, which is done by directly setting the output feature dimension as 2. Different color denotes different class and black arrow denotes the class proxy. We compare the difference between the CE loss and the HUG-MHE loss (with either independently optimized proxies or fully learnable proxies). Specifically, for the HUG-MHE loss with independently optimized proxies, we use the following form:

$$
\max_{\{x_j\}_{j=1}^n, \{w_c\}_{c=1}^C} \mathcal{L}_{\text{HUG}} := \alpha \cdot \mathcal{H}_I(\{\hat{w}_c\}_{c=1}^C) - \beta \cdot \sum_{c=1}^C \mathcal{H}_I(\{\hat{x}_i\}_{i \in A_c, SG(\hat{w}_c)})
$$

(16)

where we stop the gradient for the class proxies in the intra-class hyperspherical uniformity term.

From the results, we observe that the our HUG losses generally learns better representations than the CE loss, and moreover, HUG learns more aligned class proxy and class feature-mean than CE.

Figure 8: 2D MNIST feature visualization for the CE loss at 1,5,10,15,20 epochs (top left - top right - middle left - middle right -bottom).
Figure 9: 2D MNIST feature visualization for the HUG loss (randomly initialized and then optimized proxies) at 1,5,10,15,20 epochs (top left - top right - middle left - middle right -bottom).
Figure 10: 2D MNIST feature visualization for the HUG loss (fully learnable proxies) at 1, 5, 10, 15, 20 epochs (top left - top right - middle left - middle right - bottom).
C OTHER VARIANTS IN THE HUG FRAMEWORK

There are plenty of interesting and useful instantiations for the loss function under the HUG framework. In this section, we discuss a few highly relevant and natural ones.

C.1 PROXY-FREE HUG

We have the following general HUG objective function:

\[
\max_{\{\hat{x}_i\}_{i=1}^{n}} \mathcal{L}_{\text{HUG}} := \alpha \cdot \mathcal{HU} \left( \{\hat{\mu}_c\}_{c=1}^{C} \right) - \beta \cdot \sum_{c=1}^{C} \mathcal{HU} \left( \{\hat{x}_i\}_{i \in A_c} \right)
\]

where we can have many possible instantiations. Other than the proxy-based form proposed in the main paper, we can also have a proxy-free version:

\[
\max_{\{\hat{x}_i\}_{i=1}^{n}} \mathcal{L}_{\text{PF-HUG}} := \alpha \cdot \mathcal{HU} \left( \{\hat{x}_i\}_{i \in A_c} \right) - \beta \cdot \sum_{c=1}^{C} \mathcal{HU} \left( \{\hat{x}_i\}_{i \in A_c} \right)
\]

where \( \{\hat{x}_i\}_{i \in A_c} \) denotes a set of vectors that consist of one random sample per class. This is essentially to replace the class proxy with a random sample from this class. The proxy-free HUG loss can be used in the scenario where extremely large amount of classes exist and storing class proxies can be very expensive, or in the scenario of self-supervised contrastive learning where each instance and its augmentations are viewed as one class. A MHE-based instantiation of Eq. 18 is given by

\[
\min_{\{\hat{x}_i\}_{i=1}^{n}} \mathcal{L}_{\text{MHE-PF-HUG}} := \alpha \cdot E_{s_h} \left( \{\hat{x}_i\}_{i \in A_c} \right) - \beta \cdot \sum_{c=1}^{C} E_{s_w} \left( \{\hat{x}_i\}_{i \in A_c} \right)
\]

which can be similarly relaxed to

\[
\mathcal{L}'_{\text{MHE-PF-HUG}} = \alpha \cdot \sum_{c \neq c'} \|\hat{x}_i \in A_c - \hat{x}_j \in A_{c'}\|^{-2} + \beta' \cdot \sum_{c} \sum_{i \in A_c, j \in A_c, i \neq j} \|\hat{x}_i \in A_c - \hat{x}_j \in A_{c'}\|
\]

where \( \hat{x}_i \in A_c \) denotes a randomly selected sample from the \( c \)-th class. The first term in Eq. 20 can also be viewed as a scalable stochastic approximation to the first term in the following loss function:

\[
\mathcal{L}''_{\text{MHE-PF-HUG}} = \alpha \cdot \sum_{i \in A_c, j \in A_{c'}, c \neq c'} \|\hat{x}_i - \hat{x}_j\|^{-2} + \beta' \cdot \sum_{c} \sum_{i \in A_c, j \in A_{c'}, i \neq j} \|\hat{x}_i - \hat{x}_j\|
\]

which is typically optimized by stochastic gradients (samples come as a mini batch) in practice.

C.2 COUPLED HUG

One advantage of HUG is that it decouples intra-class variability and inter-class separability. However, coupling may also bring some benefits (e.g., robustness on hyperparameters, stability in training). To this end, we also propose a coupled loss function using the HUG framework:

\[
\max_{\{\hat{x}_i\}_{i=1}^{n}} \mathcal{L}_{\text{PF-HUG}} := \alpha \cdot \sum_{i=1}^{n} \mathcal{HU} \left( \{\hat{w}_c\}_{c=1, c \neq y_i}^{C} \hat{x}_i \right) - \beta \cdot \sum_{c=1}^{C} \mathcal{HU} \left( \{\hat{x}_i\}_{i \in A_c} \right)
\]

which can be turned into a MHE-based instantiation:

\[
\mathcal{L}'_{\text{MHE-C-HUG}} = \alpha \cdot \sum_{i=1}^{n} \sum_{c \neq y_i} \|\hat{x}_i - \hat{w}_c\|^{-2} + \beta' \cdot \sum_{c} \sum_{i \in A_c, j \in A_{c'}, i \neq j} \|\hat{x}_i - \hat{x}_j\|
\]

where the first term itself couples intra-class and inter-class hyperspherical uniformity. Although the coupled HUG drops the flexibility that the original HUG framework brings, it may introduce extra advantages (e.g., training stability).
C.3 HUG without Hyperspherical Normalization

While the CE loss does not necessarily require hyperspherical normalization for the proxies and features (but hyperspherical normalization does improve CE’s generalizability [44, 75]), we also consider the HUG framework without hyperspherical normalization here. We note that this issue remains an open challenge and we only aim to provide some simple yet natural designs.

The obvious problem to remove hyperspherical normalization is that HUG has a trivial way to decrease its loss – simply increasing the magnitude of features and proxies. A naive way to address this is to introduce magnitude penalty terms for the features and proxies. This results in

\[
\max_{\{x_i\}_{i=1}^n, \{w_c\}_{c=1}^C} \mathcal{L}_{\text{UN-P-HUG}} := \alpha \cdot \mathcal{HU}(\{w_c\}_{c=1}^C) - \beta \cdot \sum_{c=1}^C \mathcal{HU}(\{x_i\}_{i \in A_c}, w_c) - \lambda_1 \cdot \sum_{c=1}^C \|w_c - s\|^2 - \lambda_2 \cdot \sum_{i=1}^n \|x_i - s\|^2
\]

where \(s\) denotes the magnitude hyperparameter.

\[
\text{Inter-class Hyperspherical Uniformity} \quad \text{Intra-class Hyperspherical Uniformity}
\]

\[
\text{Soft Magnitude Constraint on Proxies} \quad \text{Soft Magnitude Constraint on Features}
\]
D PROOF OF THEOREM 1

We first let \( \hat{V}_C = \{ \hat{v}_1, \cdots, \hat{v}_C \} \) be an arbitrary vector configuration in \( S^{d-1} \). Then we will have that

\[
\Lambda(\hat{V}_C) := \sum_{i=1}^{C} \sum_{j=1}^{C} \| \hat{v}_i - \hat{v}_j \|^2 = \sum_{i=1}^{C} \sum_{j=1}^{C} (2 - 2\hat{v}_i \cdot \hat{v}_j) = 2C^2 - 2\left\| \sum_{i=1}^{C} \hat{v}_i \right\|^2 \leq 2C^2
\]

which holds if and only if \( \sum_{i=1}^{C} \hat{v}_i = 0 \). The vertices of a regular \((n-1)\)-simplex at the origin well satisfy this condition. With the properties of the potential function \( f \), we have that

\[
E_f(\hat{V}_C) := \sum_{i=1}^{C} \sum_{j \neq i} f(\| \hat{v}_i - \hat{v}_j \|^2) \geq C(C-1)f\left( \frac{\Lambda(\hat{V}_C)}{C(C-1)} \right) \geq C(C-1)f\left( \frac{2C}{C-1} \right)
\]

which holds true if all pairwise distance \( \| \hat{v}_i - \hat{v}_j \| \) are equal for \( i \neq j \) and the center of mass is at the origin (i.e., \( \sum_{i=1}^{C} \hat{v}_i = 0 \)). Therefore, for the vector configuration \( \hat{V}_C^* \) which contains the vertices of a regular \((C-1)\)-simplex inscribed in \( S^d \) and centered at the origin, we have that for \( 2 \leq C \leq d + 1 \)

\[
E_f(\hat{V}_C^*) = C(C-1)f\left( \frac{2C}{C-1} \right) \leq E_f(\hat{V}_C). \tag{26}
\]

If \( f \) is strictly convex and strictly decreasing, then \( E_f(\hat{V}_C) \geq C(C-1)f\left( \frac{2C}{C-1} \right) \) holds only when \( \hat{V}_C^* \) is a regular \((C-1)\)-simplex inscribed in \( S^{d-1} \) and centered at the origin. \( \blacksquare \)
E Proof of Theorem 2

This result comes as a natural conclusion from [12] where they prove that any sharp code is a minimal hyperspherical \( f \)-energy \( N \)-point configuration for any interaction potential \( f \) that is absolutely monotone on \([-1, 1]\) including all Riesz \( s \)-potentials \( f(t) = 2(t - 2t)^{-s/2} \) for \( s > 0 \).

Before we move on, we need to introduce the definition of sharp code:

**Definition 1** Let \( \hat{V}_N = \{ \hat{v}_1, \cdots, \hat{v}_N \} \) be a \( N \)-point configuration on \( \mathbb{S}^{d'} \).

- If for every \((d' + 1)\)-variate polynomial \( P \) of degree at most \( m \),
  \[
  \int_{\mathbb{S}^{d'}} P d\sigma_{d'} = \frac{1}{N} \sum_{i=1}^{N} P(\hat{v}_i)
  \]
  then \( \hat{V}_N \) is called a spherical \( m \)-design.

- If \( \hat{V}_N \) is a configuration of \( N \) distinct points such that the set of inner products between distinct points in \( \hat{V}_N \) has cardinality \( k \), then \( \hat{V}_N \) is called a spherical \( k \)-distance set.

- The configuration \( \hat{V}_N \) is a sharp code if it is both a \( k \)-distance set and a spherical \((2k - 1)\)-design.

The Cohn-Kumar Universal Optimality theorem [12] states that any sharp code is universally optimal. By universal optimality, we mean that

**Definition 2** An \( N \)-point configuration \( \hat{V}_N \) on \( \mathbb{S}^{d'} \) is called universally optimal if

\[
E_f(\hat{V}_N) := \sum_{\hat{v}_1, \hat{v}_2 \in \hat{V}_N, \hat{v}_1 \neq \hat{v}_2} f(\hat{v}_1^\top \hat{v}_2) = \min_{\hat{V}_N \subset \mathbb{S}^{d'}} E_f(\hat{V}_N)
\]

holds for any absolutely monotone function \( f : [-1, 1] \to \mathbb{R} \).

Then formally, Cohn-Kuma Universal Optimality Theorem states:

**Theorem 6** If \( \hat{V}_N \) is a sharp code on \( \mathbb{S}^{d'} \), then \( \hat{V}_N \) is universally optimal.

Because the vertices of the cross-polytope are a sharp code, then this vertex set \((2d' + 2\) points in total) is universally optimal, which implies that

\[
E_f(\hat{W}_N) = \min_{\hat{W}_N \subset \mathbb{S}^{d'}} E_f(\hat{W}_N) \tag{27}
\]

where \( \hat{W}_N \) denote the vertex set of the cross-polytope. Then we let \( s = 2 \) for the \( f \)-energy and \( d' = d - 1 \), and we prove our theorem.
F PROOF OF THEOREM 3

This theorem is in fact a well-known result (see [3, 27, 38, 63]). This general result is stated as

**Theorem 7** If \( A \subset \mathbb{R}^p \) is compact with \( \dim A > 0 \) and \( 0 < s < \dim A \), then

\[
\lim_{N \to \infty} \frac{\varepsilon_s(A, N)}{N^2} = W_s(A),
\]

where \( \varepsilon_s(A, n) := \min_{W_n \subset A} E_s(W_n) \) and \( W_s(A) \) is Wiener constant. Moreover, the equilibrium measure \( \mu_{s,A} \) on \( A \) is unique for the Riesz \( s \)-kernel when \( 0 < s < \dim A \). Finally, any sequence \( \{\hat{v}_1^N, \cdots, \hat{v}_N^N\} \) of asymptotically \( s \)-energy minimizing \( N \)-point configuration on \( A \) satisfies

\[
v(\{\hat{v}_1^N, \cdots, \hat{v}_N^N\}) \to_{\text{weak}} \mu_{s,A}, \ N \to \infty
\]

From the theorem above, with \( s = 2, \ d - 1 > s \), \( N = C \) and \( A = S^{d-1} \), we have that \( W_s(S^{d-1}) \) is a constant term, and most importantly, we have that these point sequences \( \{\hat{\mu}_1^C, \cdots, \hat{\mu}_N^C\} \) asymptotically minimizes the hyperspherical energy on \( S^{d-1} \).

Moreover, the same theorem also gives that the leading term of the minimum hyperspherical energy is of order \( O(\frac{1}{n^2}) \) as \( n \to \infty \). ■
G PROOF OF PROPOSITION 1

We show that zero-mean equal-variance Gaussian distributed vectors (after normalized to norm 1) are uniformly distributed over the unit hypersphere with Theorem 8.

**Lemma 1** Let \( \boldsymbol{x} \) be a \( n \)-dimensional random vector with distribution \( \mathcal{N}(0, 1) \) and \( \boldsymbol{U} \in \mathbb{R}^{n \times n} \) be an orthogonal matrix (\( \boldsymbol{U} \boldsymbol{U}^\top = \boldsymbol{U}^\top \boldsymbol{U} = \mathbf{I} \)). Then \( \boldsymbol{Y} = \boldsymbol{U} \boldsymbol{x} \) also has the distribution of \( \mathcal{N}(0, 1) \).

**Proof G.1** For any measurable set \( A \subset \mathbb{R}^n \), we have that

\[
P(\boldsymbol{Y} \in A) = P(\boldsymbol{X} \in \boldsymbol{U}^\top A) = \int_{\boldsymbol{U}^\top A} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \langle \boldsymbol{x}, \boldsymbol{x} \rangle} d\boldsymbol{x} = \int_{A} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \langle \boldsymbol{U} \boldsymbol{x}, \boldsymbol{U} \boldsymbol{x} \rangle} d\boldsymbol{x} = \int_{A} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \langle \boldsymbol{x}, \boldsymbol{x} \rangle} d\boldsymbol{x} \tag{28}
\]

because of orthogonality of \( \boldsymbol{U} \). Therefore the lemma holds. \( \blacksquare \)

**Theorem 8** The normalized vector of Gaussian variables is uniformly distributed on the sphere. Formally, let \( x_1, x_2, \ldots, x_n \sim \mathcal{N}(0, 1) \) and be independent. Then the vector

\[
\boldsymbol{x} = \begin{bmatrix} x_1/z, & x_2/z, & \cdots, & x_n/z \end{bmatrix}
\tag{29}
\]

follows the uniform distribution on \( S^{n-1} \), where \( z = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \) is a normalization factor.

**Proof G.2** A random variable has distribution \( \mathcal{N}(0, 1) \) if it has the density function

\[
f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x^2}. \tag{30}
\]

A \( n \)-dimensional random vector \( \boldsymbol{x} \) has distribution \( \mathcal{N}(0, 1) \) if the components are independent and have distribution \( \mathcal{N}(0, 1) \) each. Then the density of \( \boldsymbol{x} \) is given by

\[
f(x) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2} \langle \boldsymbol{x}, \boldsymbol{x} \rangle}. \tag{31}
\]

Then we use Lemma 1 about the orthogonal-invariance of the normal distribution. Because any rotation is just a multiplication with some orthogonal matrix, we know that normally distributed random vectors are invariant to rotation. As a result, generating \( \boldsymbol{x} \in \mathbb{R}^n \) with distribution \( \mathcal{N}(0, 1) \) and then projecting it onto the hypersphere \( S^{n-1} \) produces random vectors \( \boldsymbol{U} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||} \) that are uniformly distributed on the hypersphere. Therefore the theorem holds. \( \blacksquare \)

The above results indicate that as long as class proxies are initialize with zero-mean Gaussian, they are uniformly distributed over the hypersphere in a probabilistic sense. \( \blacksquare \)
H DERIVATION OF HUG SURROGATE FOR MHE AND MHS

The derivation of $\mathcal{L}_{\text{MHE-HUG}}'$ is as follows:

$$
\mathcal{L}_{\text{MHE-HUG}} := \alpha \cdot \mathbb{E}_{s_1} (\{ \hat{w}_c \}_{c=1}^C) - \beta \cdot \sum_{c=1}^C \mathbb{E}_{s_2} (\{ \hat{x}_i \}_{i \in A_c}, \hat{w}_c)
$$

$$
= \alpha \cdot \sum_{c \neq c'} \| \hat{w}_c - \hat{w}_{c'} \|^{-2} + \beta \cdot \sum_{c} \left( \sum_{i,j \in A_c, i \neq j} \| \hat{x}_i - \hat{x}_j \| + 2 \cdot \sum_{i \in A_c} \| \hat{x}_i - \hat{w}_c \| \right)
$$

$$
\leq \alpha \cdot \sum_{c \neq c'} \| \hat{w}_c - \hat{w}_{c'} \|^{-2} + \beta \cdot \sum_{c} \left( \sum_{i,j \in A_c, i \neq j} (\| \hat{x}_i - \hat{w}_c \| + \| \hat{w}_c - \hat{x}_j \|) + 2 \cdot \sum_{i \in A_c} \| \hat{x}_i - \hat{w}_c \| \right)
$$

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

The derivation of $\mathcal{L}_{\text{MHS-HUG}}'$ is as follows:

$$
\mathcal{L}_{\text{MHS-HUG}} := \alpha \cdot \theta (\{ \hat{w}_c \}_{c=1}^C) - \beta \cdot \sum_{c=1}^C \theta (\{ \hat{x}_i \}_{i \in A_c}, \hat{w}_c)
$$

$$
= \alpha \cdot \min_{c \neq c'} \| \hat{w}_c - \hat{w}_{c'} \| - \beta \cdot \sum_{c} \max_{u, v \in \{ \hat{x}_i \}_{i \in A_c}, \hat{w}_c, u \neq v} \| u - v \| \quad (33)
$$

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

Most importantly, $\sum_{c} \max_{u, v \in \{ \hat{x}_i \}_{i \in A_c}, \hat{w}_c, u \neq v} \| u - v \|$ in $\mathcal{L}_{\text{MHS-HUG}}$ and $\mathcal{L}_{\text{MHS-HUG}}'$ share the same minimizer (minimum is 0, which happens when intra-class feature collapse to its class proxy). Therefore, $\mathcal{L}_{\text{MHS-HUG}}'$ and $\mathcal{L}_{\text{MHS-HUG}}$ share the same maximizer, and $\mathcal{L}_{\text{MHS-HUG}}'$ can be viewed as a surrogate loss for $\mathcal{L}_{\text{MHS-HUG}}$. 

I Proof of Proposition 2

For notational convenience, we first define $\varepsilon_s(S^{d-1}, n) := \min_{V_n \subseteq S^{d-1}} E_s(V_n)$ and $\delta_n^s(S^{d-1}) := \max_{V_n \subseteq S^{d-1}} \theta(V_n)$. We then define that $V_n^s$ is a $s$-energy minimizing $n$-point configuration on $S^{d-1}$ if $0 < s < \infty$ (i.e., MHE configuration) and $V_n^{\infty}$ denotes a best-packing configuration on $S^{d-1}$ if $s = \infty$ (i.e., MHS configuration). Since we are considering $s > 0$, we only need to discuss the case of $K_s(\hat{v}_i, \hat{v}_j) = \rho(\hat{v}_i, \hat{v}_j)^{-s}$. Then we will have the following equation:

$$
\varepsilon_s(S^{d-1}, n) \leq E_s(\hat{V}_n^s) \geq \frac{1}{\delta_n^s(\hat{V}_n^s)} \geq \frac{1}{\delta_n^s(S^{d-1})}.
$$

(34)

Moreover, we have that

$$
\varepsilon_s(S^{d-1}, n) \leq E_s(\hat{V}_n^{\infty}) \leq \frac{1}{\delta(\hat{V}_n^{\infty})} \left( \sum_{1 \leq i \neq j \leq N} \left( \frac{\delta(\hat{V}_n^{\infty})}{\rho(\hat{v}_i^{\infty}, \hat{v}_j^{\infty})} \right)^s \right)^{\frac{1}{s}} \leq \frac{1}{\delta(\hat{V}_n^{\infty}) (n(n - 1))^s}.
$$

(35)

Therefore, we will end up with

$$
\lim_{s \to \infty} \sup_{\varepsilon_s(S^{d-1}, n)} \leq \frac{1}{\delta(\hat{V}_n^{\infty})} = \frac{1}{\delta_n^s(S^{d-1})}.
$$

(36)

Then we take both Eq. 34 and Eq. 36 into consideration and have that

$$
\lim_{s \to \infty} \varepsilon_s(S^{d-1}, n) \leq \frac{1}{\delta_n^s(S^{d-1})}.
$$

(37)

which concludes the proof. ■
J  Proof of Proposition 3

We write down the formulation of the HUG objectives (with MHE):

\[
\min_{\{\hat{x}_i\}_{i=1}^n, \{\hat{w}_c\}_{c=1}^C} \mathcal{L}_{\text{MHE-HUG}} := \alpha \cdot E_{sb}(\{\hat{w}_c\}_{c=1}^C) - \beta \cdot \sum_{c=1}^C E_{sw}(\{\hat{x}_i\}_{i \in A_c}, \hat{w}_c) \\
= \alpha \cdot \sum_{c \neq c'} \|\hat{w}_c - \hat{w}_{c'}\|^{-2} + \beta \cdot \sum_c \left( \sum_{i,j \in A_c, i \neq j} \|\hat{x}_i - \hat{x}_j\| + 2 \cdot \sum_{i \in A_c} \|\hat{x}_i - \hat{w}_c\| \right) \\
\]

(38)

\[
\min_{\{\hat{x}_i\}_{i=1}^n, \{\hat{w}_c\}_{c=1}^C} \mathcal{L}'_{\text{MHE-HUG}} = \alpha \cdot \sum_{c \neq c'} \|\hat{w}_c - \hat{w}_{c'}\|^{-2} + \beta' \cdot \sum_c \sum_{i \in A_c} \|\hat{x}_i - \hat{w}_c\| \\
\]

(39)

For both objectives, we can see that the minimizer of the second term (i.e., the intra-class variability term) is all intra-class feature collapse to their class proxy and therefore the second term achieves the global minimum 0.

For the first term of both objectives, the global minimizer can be obtain directly from Theorem 1, Theorem 2 and Theorem 3. It is easy to see that the global minimizer of the inter-class separability term and the intra-class variability term does not contradict with each other and can be achieved simultaneously.
K Proof of Proposition 4

\[
\sum_{i=1}^{n} \log(1 + \sum_{j=1}^{C} \exp(\langle w_j, x_i \rangle - \langle w_{y_i}, x_i \rangle)) \\
\geq \sum_{i=1}^{n} \sum_{j=1}^{C} \log(1 + \exp(\langle w_j, x_i \rangle - \langle w_{y_i}, x_i \rangle)) \\
\geq \sum_{i=1}^{n} \sum_{j=1}^{C} (\langle w_j, x_i \rangle - \langle w_{y_i}, x_i \rangle) \\
= \sum_{i=1}^{n} \sum_{j \neq y_i}^{C} (\langle w_j, x_i \rangle) - (C - 1) \sum_{i=1}^{n} \langle w_{y_i}, x_i \rangle \\
\]  

\[Q_1: \text{Coupling IS and IV} \quad Q_2: \text{Inter-class Variability} \]

\[
\sum_{i=1}^{n} \log(1 + \sum_{j=1 \neq y_i}^{C} \exp(\langle w_j, x_i \rangle - \langle w_{y_i}, x_i \rangle)) \\
\leq \log(1 + \sum_{i=1}^{n} \sum_{j=1 \neq y_i}^{C} \exp(\langle w_j, x_i \rangle - \langle w_{y_i}, x_i \rangle)) \\
\leq \log(1 + \sum_{i=1}^{n} \sum_{j=1 \neq y_i}^{C} (\exp(\langle w_j, x_i \rangle) + \exp(-\langle w_{y_i}, x_i \rangle))) \\
= \log \left(1 + \sum_{i=1}^{n} \sum_{j \neq y_i}^{C} \exp(\langle w_j, x_i \rangle) + (C - 1) \sum_{i=1}^{n} \exp(-\langle w_{y_i}, x_i \rangle) \right) \\
\]  

\[Q_1: \text{Coupling IS and IV} \quad Q_2: \text{Inter-class Variability} \]
I. DERIVATION OF CE’S LOWER BOUND

The derivation is actually very simple and this result is originally given by [4]. We find that it naturally matches the intuition behind HUG. For our paper to be self-contained, we briefly give the simple derivation below. For the details, please refer to Proposition 1 in [4].

We start by rewriting the CE loss as

\[
L_{CE} = -\sum_{i=1}^{n} \langle w_{y_i}, x_i \rangle + \frac{\lambda n}{2} \sum_{c=1}^{C} \langle w_c, w_c \rangle + \sum_{i=1}^{n} \log \sum_{c=1}^{C} \exp(\langle w_c, x_i \rangle) - \frac{\lambda n}{2} \sum_{c=1}^{C} \langle w_c, w_c \rangle \tag{42}
\]

where \(\lambda\) can be chosen such that both \(Q_1(w)\) and \(Q_2(w)\) become convex functions with respect to \(w\). Taking advantage of the convexity, we can separately set the gradient of \(Q_1(w)\) and \(Q_2(w)\) with respect to \(w\) as 0 and compute their minima. Specifically, we end up with

\[
Q_1(w) \geq Q_1(w^*_Q_1) = -\frac{1}{2\lambda n} \sum_{i=1}^{n} \sum_{j \in A_{y_i}} \langle x_i, x_j \rangle, \tag{43}
\]

\[
Q_2(w) \geq Q_2(w^*_Q_2) = \sum_{i=1}^{n} \log \sum_{c=1}^{C} \exp \left( \frac{1}{\lambda n} \sum_{j=1}^{n} l_{ic} \langle x_i, x_j \rangle \right) - \frac{n}{2\lambda} \sum_{c=1}^{C} \left\| \frac{1}{n} \sum_{i=1}^{n} l_{ic} x_i \right\|^2, \tag{44}
\]

where \(l_{ic} = \frac{\exp(\langle w_c, x_i \rangle)}{\sum_{j=1}^{n} \exp(\langle w_j, x_i \rangle)}\) denotes the softmax confidence. Combining the two lower bounds above, we can have that

\[
L_{CE} \geq Q_1(w^*_Q_1) + Q_2(w^*_Q_2)
= \sum_{i=1}^{n} \log \sum_{c=1}^{C} \exp \left( \frac{1}{\lambda n} \sum_{j=1}^{n} l_{jc} \langle x_i, x_j \rangle \right) - \frac{1}{2\lambda} \sum_{c=1}^{C} \left\| \frac{1}{n} \sum_{i=1}^{n} l_{ic} x_i \right\|^2 - \frac{1}{2\lambda n} \sum_{i=1}^{n} \sum_{j \in A_{y_i}} \langle x_i, x_j \rangle \tag{45}
\]

where the first two terms encourage larger inter-class hyperspherical uniformity, and the last term promotes smaller intra-class hyperspherical uniformity.
M PROOF OF THEOREM 5

This theorem follows naturally from the main result in [50]. [50] has proved that the minimizer of a simplified form of the cross-entropy loss is the simplex ETF when $2 \leq C \leq d + 1$ and the minimizer also asymptotically converges to uniform measure on the hypersphere. More formally, we have

**Theorem 9 ([50])** Consider the following variational problem

$$
\min_{\mathbf{u}} \mathcal{L}_\alpha(\mathbf{u}) := \sum_{i=1}^{n} \log \left( \frac{\sum_{j=1}^{n} \exp (\langle \mathbf{u}_j, \mathbf{u}_i \rangle)}{\exp (\langle \mathbf{u}_i, \mathbf{u}_i \rangle)} \right) \tag{46}
$$

s.t. $\mathbf{u}_i \in \mathbb{R}^d, \|\mathbf{u}_i\| = 1, \forall i$

Let $\mu_n$ be the probability measure on $\mathbb{S}^d$ generated by a minimizer

$$
\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{u}_i}, \tag{47}
$$

then for any $\alpha > 0$, $\mu_n$ converges weakly to the uniform measure on $\mathbb{S}^{d-1}$ as $n \to \infty$.

From Theorem 3, we know that HUG with specific potential energy also converges to the uniform measure on $\mathbb{S}^{d-1}$. Combining the results above, we can conclude that HUG and CE share the same minimizer. ■
E XPERIMENTAL DETAILS

General settings. For MHE-HUG and MHS-HUG, $\alpha$ and $\beta$ are set as 0.15 and 0.015, respectively. For MGD-HUG, $\alpha$ and $\beta$ are set as 0.15 and 0.03, respectively. We train the model for 200 epochs with 512 batchsize for both the cross-entropy (CE) loss and HUG. We use the stochastic gradient descent with momentum 0.9 and weight decay $2 \times 10^{-4}$. The initial learning rate is set as 0.1 for both CIFAR-100 and CIFAR-10 and is divided by 10 at 60, 120, 180 epoch. For the general classification experiments, we use multiple architectures, including ResNet-18, VGG16 and DenseNet121. We use the simple data augmentation: 4 pixels are padded on each side, and image is randomly cropped.

Long-tailed recognition. We follow LDAM [6] to obtain imbalanced CIFAR-10 and CIFAR-100 datasets with different imbalanced ratio. Following LDAM, we use ResNet-32 as our base network. The other setting is the same as our general setting.

Continual learning. We follow DER [5] to construct our continual learning experiments. We split both the CIFAR-10 and CIFAR-100 training set into 5 tasks. Each task has 2 classes and 20 classes for CIFAR-10 and CIFAR-100, respectively. The training batchsize is set as 64, where there are 32 incoming samples and 32 replayed samples. Different size of memory buffer is also studied.

Adversarial robustness. For the experiments of adversarial robustness, we first obtain the model trained with CE and HUG. With the information of the attacked model, PGD [51] generates some adversarial examples to mislead the attacked model. The test accuracy in the experiments of adversarial robustness shows the accuracy of the perturbed samples.

Visualizing loss landscape. We perturb neuron weights to visualize the loss landscape, as proposed in [40]. For details, we perturb the model weight with 400 interpolation points in two random vectors around the current model weight minima. The visualization method is also the same as [47].
O ADDITIONAL EXPERIMENTAL RESULTS

Training convergence. We observe the training convergence of HUG on CIFAR-10 and CIFAR-100. Both the evaluation accuracy and the training loss, including the overall losses, the intra-class loss and the inter-class loss, are shown in Figure 11. For both the CIFAR-10 and CIFAR-100, the inter-class uniformity loss remains relatively small, which is consistent with the empirical finding in [41, 47]. Moreover, we find that the intra-class uniformity loss (i.e., intra-class variability) dominates the overall loss on CIFAR-100 dataset and it is relatively difficult to optimize when the class number becomes large.

2D loss contour. We also utilize the method in [40] to visualize the 2D loss landscape, which is more easy to visualize the flatness of the loss landscape. As shown in Figure 12, the 2D loss landscape of our HUG loss is flatter than the widely used CE loss, showing that HUG yields a flat minima which may have better generalization ability.

The ablation of $\alpha$ and $\beta$. In our HUG framework, we introduce two scaling hyperparameters, $\alpha$ for the inter-class hyperspherical uniformity, $\beta$ for the intra-class hyperspherical uniformity. We investigate the effect of the two hyperparameters for the model performance. As shown in Table 8, HUG is not sensitive to $\alpha$, as the inter-class hyperspherical uniformity is always easy to optimize. HUG is also not sensitive to $\beta$ in a wide range. The ablations are conducted on CIFAR-100. $\alpha$ is set as 0.15 when we perform ablation on $\beta$. $\beta$ is set as 0.015 when doing ablation on $\alpha$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.0003</th>
<th>0.0015</th>
<th>0.015</th>
<th>0.05</th>
<th>0.15</th>
<th>0.5</th>
<th>1.5</th>
<th>5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>76.31</td>
<td>75.99</td>
<td>76.28</td>
<td>76.16</td>
<td>76.48</td>
<td>76.32</td>
<td>76.1</td>
<td>76.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>0.005</th>
<th>0.015</th>
<th>0.05</th>
<th>0.15</th>
<th>0.3</th>
<th>0.5</th>
<th>1.5</th>
<th>5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>74.15</td>
<td>76.48</td>
<td>76.12</td>
<td>75.87</td>
<td>75.59</td>
<td>75.24</td>
<td>74.81</td>
<td>74.00</td>
</tr>
</tbody>
</table>

Table 8: Effect of hyperparameters $\alpha$ and $\beta$. 