Appendix

A Proof of Proposition 1

We first define that $\hat{\boldsymbol{W}}_n^s$ is a s-energy minimizing n-point configuration on $\mathbb{S}^{d-1}$ uf $0 < s < \infty$ (*i.e.*, MHE configuration) and $\hat{\boldsymbol{W}}_n^\infty$ denotes a best-packing configuration on $\mathbb{S}^{d-1}$ if $s = \infty$ (*i.e.*, MHS configuration). Since we are considering $s > 0$, we only need to discuss the case of $K_s(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j) = \rho(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j)^{-s}$. Then we will have the following equation:

$$\varepsilon_s(\mathbb{S}^{d-1}, n)^{\frac{1}{s}} = E_s(\hat{\boldsymbol{W}}_n^s)^{\frac{1}{s}} \geq \frac{1}{\delta_n^\rho(\hat{\boldsymbol{W}}_n^s)} \geq \frac{1}{\delta_n^\rho(\mathbb{S}^{d-1})}. \tag{18}$$

Moreover, we have that

$$\begin{aligned} \varepsilon_s(\mathbb{S}^{d-1}, n)^{\frac{1}{s}} &\leq E_s(\hat{\boldsymbol{W}}_n^\infty)^{\frac{1}{s}} \\ &= \frac{1}{\delta^\rho(\hat{\boldsymbol{W}}_n^\infty)} \Bigg(\sum_{1 \leq i \neq j \leq N} \Big(\frac{\delta^\rho(\hat{\boldsymbol{W}}_n^\infty)}{\rho(\hat{\boldsymbol{w}}_i^\infty, \hat{\boldsymbol{w}}_j^\infty)} \Big)^s \Bigg)^{\frac{1}{s}} \\ &\leq \frac{1}{\delta^\rho(\hat{\boldsymbol{W}}_n^\infty)} \big(n(n-1) \big)^{\frac{1}{s}} \end{aligned} \tag{19}$$

Therefore, we will end up with

$$\lim_{s \to \infty} \sup \varepsilon_s(\mathbb{S}^{d-1}, n)^{\frac{1}{s}} \leq \frac{1}{\delta^\rho(\hat{\boldsymbol{W}}_n^\infty)} = \frac{1}{\delta_n^\rho(\mathbb{S}^{d-1})}. \tag{20}$$

Then we take both Eq. (18) and Eq. (20) into consideration and have that

$$\lim_{s \to \infty} \varepsilon_s(\mathbb{S}^{d-1}, n)^{\frac{1}{s}} = \frac{1}{\delta_n^\rho(\mathbb{S}^{d-1})} \tag{21}$$

which concludes the proof. □

B Proof of Proposition 2

We first choose $\epsilon > 0$ and let $\hat{\boldsymbol{W}}_{n+1} = \{\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2, \cdots, \hat{\boldsymbol{w}}_{n+1}\} \subset \mathbb{S}^{d-1}$ be a configuration such that

$$\varepsilon_s(\mathbb{S}^{d-1}, n+1) + \epsilon > E_s(\hat{\boldsymbol{W}}_{n+1}). \tag{22}$$

Then we have for every $i \in [1, n+1]$ and $\boldsymbol{v} \in \mathbb{S}^{d-1}$ that

$$\begin{aligned} E_s\big((\hat{\boldsymbol{W}}_{n+1} \backslash \{\hat{\boldsymbol{w}}_i\}) \cup \{\boldsymbol{v}\}\big) &= E_s(\hat{\boldsymbol{W}}_{n+1} \backslash \{\hat{\boldsymbol{w}}_i\}) + 2 \sum_{j:j \neq i} K_s(\boldsymbol{v}, \hat{\boldsymbol{w}}_j) \\ &\geq \varepsilon_s(\mathbb{S}^{d-1}, n+1) \\ &> E_s(\hat{\boldsymbol{W}}_{n+1}) - \epsilon \\ &= E_s(\hat{\boldsymbol{W}}_{n+1} \backslash \{\boldsymbol{v}\}) + 2 \sum_{j:j \neq i} K_s(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j) - \epsilon \end{aligned} \tag{23}$$

which leads to

$$\min_{\boldsymbol{v} \in \mathbb{S}^{d-1}} 2 \sum_{j:j \neq i} K_s(\boldsymbol{v}, \hat{\boldsymbol{w}}_j) \geq 2 \sum_{j:j \neq i} K_s(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j) - \epsilon \tag{24}$$

Therefore, for a fixed i, we have that

$$\begin{aligned} \mathcal{P}_s(\mathbb{S}^{d-1}, n) &\geq P_s(\hat{\boldsymbol{W}}_{n+1} \backslash \{\hat{\boldsymbol{w}}_i\}) \\ &= \min_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \sum_{j:j \neq i} K_s(\boldsymbol{v}, \hat{\boldsymbol{w}}_j) \\ &\geq \sum_{j:j \neq i} K_s(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j) - \frac{\epsilon}{2} \end{aligned} \tag{25}$$

Then we average the above inequalities for $i = 1, \cdots, n+1$ and obtain

$$\begin{aligned}\mathcal{P}_s(\mathbb{S}^{d-1}, n) &\geq \frac{1}{n+1}\sum_{i=1}^{n+1}\sum_{j:j\neq i} K_s(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j) - \frac{\epsilon}{2} \\ &\geq \frac{\varepsilon_s(\mathbb{S}^{d-1}, n+1)}{n+1} - \frac{\epsilon}{2}\end{aligned} \tag{26}$$

By letting ϵ approach to zero, we have that

$$\mathcal{P}_s(\mathbb{S}^{d-1}, n) \geq \frac{\varepsilon_s(\mathbb{S}^{d-1}, n+1)}{n+1} \tag{27}$$

Moreover, it is also easy to verify another inequality:

$$\frac{\varepsilon_s(\mathbb{S}^{d-1}, n+1)}{n+1} \geq \frac{\varepsilon_s(\mathbb{S}^{d-1}, n)}{n-1} \tag{28}$$

Therefore, we conclude the proof. □

C Proof of Proposition 3

Given that $s = -2$, we first have that

$$\begin{aligned}P_{-2}(\hat{\boldsymbol{W}}_n) &= \min_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \left(-\sum_{i=1}^{n} \|\boldsymbol{v} - \hat{\boldsymbol{w}}_i\|^2 \right) \\ &= \min_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \sum_{i=1}^{n} (2\boldsymbol{v}\cdot\hat{\boldsymbol{w}}_i - 2) \\ &= \min_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \left(2\boldsymbol{v}\cdot\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i - 2n \right).\end{aligned} \tag{29}$$

If $\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i = \boldsymbol{0}$, we will have that $P_{-2}(\hat{\boldsymbol{W}}_n) = -2n$. If $\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i \neq \boldsymbol{0}$, then we have that

$$\begin{aligned}P_{-2}(\hat{\boldsymbol{W}}_n) &\leq -2\frac{\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i}{\|\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i\|}\cdot\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i - 2n \\ &= -2\left\|\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i\right\| - 2n \\ &< -2n\end{aligned} \tag{30}$$

Therefore, $\hat{\boldsymbol{W}}_n$ is optimal if and only if $\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i = \boldsymbol{0}$ □

D Proof of Proposition 4

For any n-point configuration $\hat{\boldsymbol{W}}_n \subset \mathbb{S}^{d-1}$, we have that

$$\begin{aligned}P_s(\hat{\boldsymbol{W}}_n) &= \min_{\boldsymbol{v}\in\mathbb{S}^{d-1}} \sum_{\boldsymbol{u}\in\hat{\boldsymbol{W}}_n} \frac{1}{\rho(\boldsymbol{v},\boldsymbol{u})^s} \\ &\geq \frac{1}{\alpha(\hat{\boldsymbol{W}}_n)^s}\end{aligned} \tag{31}$$

which leads to

$$\begin{aligned}\left(\mathcal{P}_s(\mathbb{S}^{d-1}, n)\right)^{\frac{1}{s}} &= \max_{\hat{\boldsymbol{W}}_n\subset\mathbb{S}^{d-1}} P_s(\hat{\boldsymbol{W}}_n)^{\frac{1}{s}} \\ &\geq \max_{\hat{\boldsymbol{W}}_n\subset\mathbb{S}^{d-1}} \frac{1}{\alpha(\hat{\boldsymbol{W}}_n)} \\ &= \frac{1}{\eta_n^{\rho}(\mathbb{S}^{d-1})}.\end{aligned} \tag{32}$$

Therefore, we have that

$$\lim_{s\to\infty}\inf\left(\mathcal{P}_s(\mathbb{S}^{d-1},n)\right)^{\frac{1}{s}}\geq\frac{1}{\eta_n^\rho(\mathbb{S}^{d-1})} \tag{33}$$

On the other hand, we have that

$$\begin{aligned}P_s(\hat{\boldsymbol{W}}_n)&=\min_{\boldsymbol{v}\in\mathbb{S}^{d-1}}\sum_{\boldsymbol{u}\in\hat{\boldsymbol{W}}_n}\frac{1}{\rho(\boldsymbol{v},\boldsymbol{u})^s}\\&\leq\frac{n}{\alpha(\hat{\boldsymbol{W}}_n)}\\&\leq\frac{n}{\left(\eta_n^\rho(\mathbb{S}^{d-1})\right)^s}\end{aligned} \tag{34}$$

Therefore, we end up with

$$\lim_{s\to\infty}\sup\left(\mathcal{P}_s(\mathbb{S}^{d-1},n)\right)^{\frac{1}{s}}\leq\lim_{s\to\infty}\frac{n^{\frac{1}{s}}}{\eta_n^\rho(\mathbb{S}^{d-1})}=\frac{1}{\eta_n^\rho(\mathbb{S}^{d-1})} \tag{35}$$

Combining with Eq. (33), we have that

$$\lim_{s\to\infty}\left(\mathcal{P}_s(\mathbb{S}^{d-1},n)\right)^{\frac{1}{s}}\geq\frac{1}{\eta_n^\rho(\mathbb{S}^{d-1})} \tag{36}$$

which concludes the proof. □

E Proof of Proposition 5

We first define the order samples on $\mathbb{S}^1$. We denote the samples on $\mathbb{S}^1$ as θ_i. The angles are ordered such that $\theta_{i+1}<\theta_i,\forall i$. Then we define the angle gap as follows:

$$\begin{aligned}d_i&:=\theta_{i+1}-\theta_i,\quad i=1,2,\cdots,n-1\\d_n&:=2\pi-(\theta_n-\theta_1)\end{aligned} \tag{37}$$

The test statistic of range test is written as

$$T_n:=2\pi-\max_i d_i \tag{38}$$

which rejects $\mathcal{H}_0$ for small values. Maximizing T_n with respect to the samples on $\mathbb{S}^1$ is equivalent to the following objective:

$$\max_{\{\theta_1,\cdots,\theta_n\}}T_n\Leftrightarrow\min_{\{\theta_1,\cdots,\theta_n\}}\max_i d_i \tag{39}$$

which is to minimize the largest neighbor angle gap. It is easy to verify that the optimum happens when the n angle gaps are equally divided the unit circle $\mathbb{S}^1$.

For MHS on $\mathbb{S}^1$, the optimization is as follows:

$$\max_{\{\theta_1,\cdots,\theta_n\}}\min_{i\neq j}\rho(\theta_i,\theta_j) \tag{40}$$

which is to maximize the smallest pairwise angles (*i.e.*, the smallest neighbor angle gap on $\mathbb{S}^1$). The optimum is attained when $\{\theta_1,\cdots,\theta_n\}$ are equally divided the unit circle $\mathbb{S}^1$, which is equivalent to maximizing T_n with respect to the samples on $\mathbb{S}^1$.

For MHC on $\mathbb{S}^1$, the optimization is as follows:

$$\min_{\{\theta_1,\cdots,\theta_n\}}\max_{v\in[0,2\pi)}\min_i\rho(v,\theta_i). \tag{41}$$

The optimum of $\max_{v\in[0,2\pi)}\min_i\rho(v,\theta_i)$ is attained when v lies on the middle point of the largest angle gap. Therefore, the optimum of MHC on $\mathbb{S}^1$ is achieved when $\{\theta_1,\cdots,\theta_n\}$ are equally divided the unit circle $\mathbb{S}^1$, which is also equivalent to maximizing T_n with respect to the samples on $\mathbb{S}^1$. □

F Proof of Theorem 2

We first let $\hat{\boldsymbol{W}}_n = \{\hat{\boldsymbol{w}}_1, \cdots, \hat{\boldsymbol{w}}_n\}$ be an arbitrary vector configuration in $\mathbb{S}^d$. We then have that

$$\begin{aligned}\Lambda(\hat{\boldsymbol{W}}_n) :=& \sum_{i=1}^{n}\sum_{j=1}^{n} \|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^2 \\ =& \sum_{i=1}^{n}\sum_{j=1}^{n} (2 - 2\hat{\boldsymbol{w}}_i \cdot \hat{\boldsymbol{w}}_j) \\ =& 2n^2 - 2\left\|\sum_{i=1}^{n} \hat{\boldsymbol{w}}_i\right\|^2 \\ \leq& 2n^2\end{aligned} \tag{42}$$

which holds if and only if $\sum_{i=1}^{n} \hat{\boldsymbol{w}}_i = 0$. 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

$$\begin{aligned}E_f(\hat{\boldsymbol{W}}_n) :=& \sum_{i=1}^{n}\sum_{j:j\neq i} f\big(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^2\big) \\ \geq& n(n-1) f\left(\frac{\Lambda(\hat{\boldsymbol{W}}_n)}{n(n-1)}\right) \\ \geq& n(n-1) f\left(\frac{2n}{n-1}\right)\end{aligned} \tag{43}$$

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

$$\begin{aligned}E_f(\hat{\boldsymbol{W}}_n^*) &= n(n-1) f\left(\frac{2n}{n-1}\right) \\ &\leq E_f(\hat{\boldsymbol{W}}_n).\end{aligned} \tag{44}$$

If f is strictly convex and strictly decreasing, then $E_f(\hat{\boldsymbol{W}}_n) \geq n(n-1) f(\frac{2n}{n-1})$ holds only when $\hat{\boldsymbol{W}}_n^*$ is a regular $(n-1)$-simplex inscribed in $\mathbb{S}^d$ and centered at the origin. □

G Proof of Theorem 4

Let $\hat{\boldsymbol{w}}_1^*, \hat{\boldsymbol{w}}_2^*, \cdots, \hat{\boldsymbol{w}}_n^*$ be the points in the MHE solution $\hat{\boldsymbol{W}}_n^*$. Without loss of generality, we denote the indices k and l such that $\vartheta(\hat{\boldsymbol{W}}_n^*) = \|\hat{\boldsymbol{w}}_k^* - \hat{\boldsymbol{w}}_l^*\|_2$. We also define $\boldsymbol{z} := (1 + n^{-\frac{1}{d-1}})\hat{\boldsymbol{w}}_k^*$. We first introduce the following fact about closed convex sets:

Proposition 6. *Let $\boldsymbol{K} \subset \mathbb{R}^p$ be a closed convex set. Then for every $\boldsymbol{x} \in \mathbb{R}^p$, there is a unique point $\boldsymbol{y}_x$ in $\boldsymbol{K}$ closest to $\boldsymbol{x}$. Furthermore, for any $\boldsymbol{z} \in \boldsymbol{K}$, we have $\|\boldsymbol{y}_x\boldsymbol{z}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{z}\|_2$, where the equality holds if and only if $\boldsymbol{x} \in \boldsymbol{K}$.*

Because the unit hyperball $B(\boldsymbol{0}, 1)$ is convex and $\hat{\boldsymbol{w}}_k^*$ is the point in $B(\boldsymbol{0}, 1)$ closest to $\boldsymbol{z}$, for $1 \leq j \leq n$ we have the following inequality based on this proposition above:

$$\|\hat{\boldsymbol{w}}_k^* - \hat{\boldsymbol{w}}_j^*\|_2 \leq \|\boldsymbol{z} - \hat{\boldsymbol{w}}_j^*\|_2, \tag{45}$$

where $1 \leq j \leq n$. Before we proceed, we need to introduce the following lemmas:

Lemma 1 ([40])**.** *If $0 < s < d-1$ and $\hat{\boldsymbol{W}}_n^* = \{\hat{\boldsymbol{w}}_1^*, \cdots, \hat{\boldsymbol{w}}_n^*\}$ is a MHE solution on $\mathbb{S}^{d-1}$, then for $i = 1, 2, \cdots, n$, we have that*

$$\frac{1}{n-1} \sum_{j:j\neq i} \frac{1}{\|\hat{\boldsymbol{w}}_i^* - \hat{\boldsymbol{w}}_j^*\|_2^s} \leq I_s[\sigma_{d-1}] \tag{46}$$

where $I_s[\mu] = \int\int \frac{1}{\|\boldsymbol{x}-\boldsymbol{y}\|_2^s} d\mu(\boldsymbol{x}) d\mu(\boldsymbol{y})$ and σ_{d-1} is the normalized probability surface area measure on $\mathbb{S}^{d-1}$.

Lemma 2 ([40])**.** *We assume $d-2 \leq s < d-1$, and then there is a constant $\theta_{s,d}$ and a positive integer m such that for every $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\|_2 = 1 + n^{-\frac{1}{d-1}}$ and any optimal MHE solution $\hat{\boldsymbol{W}}_n^*$ on $\mathbb{S}^{d-1}$, we have*

$$U_s(\boldsymbol{x}; \hat{\boldsymbol{W}}_n^*) \geq I_s[\sigma_{d-1}] - \theta_{s,d} \cdot n^{-1+\frac{s}{d-1}} \tag{47}$$

where $n > m$ and $U_s(\boldsymbol{x}; \hat{\boldsymbol{W}}_n^*) := \frac{1}{n} \sum_{\boldsymbol{y} \in \hat{\boldsymbol{W}}_n^*} \frac{1}{\|\boldsymbol{x}-\boldsymbol{y}\|_2^s}$ for $s > 0$.

Using Lemma 1 above, we obtain that

$$\begin{aligned} I_s[\sigma_{d-1}] - \frac{1}{n\vartheta(\hat{\boldsymbol{W}}_n^*)^s} &\geq \frac{1}{n}\Big(\sum_{j:j\neq k} \frac{1}{\|\hat{\boldsymbol{w}}_k^* - \hat{\boldsymbol{w}}_j^*\|_2^s} - \frac{1}{\|\hat{\boldsymbol{w}}_k^* - \hat{\boldsymbol{w}}_l^*\|_2^s} \Big) \\ &= \frac{1}{n} \sum_{j:j\neq k,l} \frac{1}{\|\hat{\boldsymbol{w}}_k^* - \hat{\boldsymbol{w}}_j^*\|_2^s} \\ &\geq \frac{1}{n} \sum_{j:j\neq k,l} \frac{1}{\|\boldsymbol{z} - \hat{\boldsymbol{w}}_j^*\|_2^s} \\ &= U_s(\boldsymbol{z}; \hat{\boldsymbol{W}}_n^*) - \frac{1}{n}\Big(\frac{1}{\|\boldsymbol{z} - \hat{\boldsymbol{w}}_k^*\|_2^s} + \frac{1}{\|\boldsymbol{z} - \hat{\boldsymbol{w}}_l^*\|_2^s} \Big). \end{aligned} \tag{48}$$

Because of $n^{-\frac{1}{d-1}} = \|\boldsymbol{z} - \hat{\boldsymbol{w}}_k^*\|_2^s \leq \|\boldsymbol{z} - \hat{\boldsymbol{w}}_l^*\|_2^s$, we have that

$$I_s[\sigma_{d-1}] - \frac{1}{n\vartheta(\hat{\boldsymbol{W}}_n^*)^s} \geq U_s(\boldsymbol{z}; \hat{\boldsymbol{W}}_n^*) - 2n^{-1+\frac{s}{d-1}} \tag{49}$$

Then according to Lemma 2, we have that

$$I_s[\sigma_{d-1}] - \frac{1}{n\vartheta(\hat{\boldsymbol{W}}_n^*)^s} \geq I_s[\sigma_{d-1}] - (\theta_{s,d} + 2) \cdot n^{-1+\frac{s}{d-1}} \tag{50}$$

which concludes that $\vartheta(\hat{\boldsymbol{W}}_n^*) \geq \lambda_{s,d} \cdot n^{-\frac{1}{d-1}}$ where we define that $\lambda_{s,d} = (\theta_{s,d} + 2)^{-\frac{1}{s}}$. Note that, the extended and generalized version of this result can be found in [21, 12, 40]. □

H Proof of Theorem 5

The theorem comes directly from the result in [43] that every asymptotically optimal MHS sequence $\{\hat{\boldsymbol{W}}_n^*\}_{n=2}^{\infty}$ of n-point configurations on $\mathbb{S}^{d-1}$ is asymptotically optimal MHE solution for any $0 < s < d-1$.

I Proof of Theorem 6

We first introduce the following lemma as the characterization of a unit vector that is uniformly distributed on the unit hypersphere $\mathbb{S}^{d-1}$.

Lemma 3 ([59]). *Let $\boldsymbol{v}$ be a random vector that is uniformly distributed on the unit hypersphere $\mathbb{S}^{d-1}$. Then $\boldsymbol{v}$ has the same distribution as the following:*

$$\left\{\frac{u_1}{\sqrt{\sum_{i=1}^d u_i^2}}, \frac{u_2}{\sqrt{\sum_{i=1}^d u_i^2}}, \cdots, \frac{u_d}{\sqrt{\sum_{i=1}^d u_i^2}}\right\} \tag{51}$$

where $u_1, u_2, \cdots, u_d$ are i.i.d. *standard normal random variables.*

Proof. The lemma follows naturally from the fact that the Gaussian vector $\{u_i\}_{i=1}^d$ is rotationally invariant. □

Then we consider a random matrix $\tilde{\boldsymbol{W}} = \{\tilde{\boldsymbol{v}}_1, \cdots, \tilde{\boldsymbol{v}}_n\}$ where $\tilde{\boldsymbol{v}}_i$ follows the same distribution of $\{u_1, \cdots, u_d\}$. Therefore, it is also equivalent to a random matrix with each element distributed normally. For such a matrix $\tilde{\boldsymbol{W}}$, we have from [72] that

$$\begin{aligned}\lim_{n\to\infty} \sigma_{\max}(\tilde{\boldsymbol{W}}) &= \sqrt{d} + \sqrt{\lambda d} \\ \lim_{n\to\infty} \sigma_{\min}(\tilde{\boldsymbol{W}}) &= \sqrt{d} - \sqrt{\lambda d}\end{aligned} \tag{52}$$

where $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ denote the largest and the smallest singular value, respectively.

Then we write the matrix $\boldsymbol{W}$ as follows:

$$\begin{aligned}\boldsymbol{W} &= \tilde{\boldsymbol{W}} \cdot \boldsymbol{Q} \\ &= \tilde{\boldsymbol{W}} \cdot \begin{bmatrix} \frac{1}{\|\tilde{\boldsymbol{v}}_1\|_2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\|\tilde{\boldsymbol{v}}_2\|_2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{\|\tilde{\boldsymbol{v}}_n\|_2} \end{bmatrix}\end{aligned} \tag{53}$$

which leads to

$$\begin{aligned}\lim_{n\to\infty} \sigma_{\max}(\boldsymbol{W}) &= \lim_{n\to\infty} \sigma_{\max}(\tilde{\boldsymbol{W}} \cdot \boldsymbol{Q}) \\ \lim_{n\to\infty} \sigma_{\min}(\boldsymbol{W}) &= \lim_{n\to\infty} \sigma_{\min}(\tilde{\boldsymbol{W}} \cdot \boldsymbol{Q})\end{aligned}. \tag{54}$$

We fist assume that for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_n(\boldsymbol{A})$. Then we introduce the following inequalities for eigenvalues:

Lemma 4 ([54]). *Let $\boldsymbol{G}, \boldsymbol{H} \in \mathbb{R}^{n\times n}$ be positive semi-definite symmetric, and let $1 \leq i_1 < \cdots < i_k \leq n$. Then we have that*

$$\prod_{t=1}^k \lambda_{i_t}(\boldsymbol{G}\boldsymbol{H}) \leq \prod_{t=1}^k \lambda_{i_t}(\boldsymbol{G})\lambda_t(\boldsymbol{H}) \tag{55}$$

and

$$\prod_{t=1}^k \lambda_{i_t}(\boldsymbol{G}\boldsymbol{H}) \geq \prod_{t=1}^k \lambda_{i_t}(\boldsymbol{G})\lambda_{n-t+1}(\boldsymbol{H}) \tag{56}$$

where λ_i denotes the i-th largest eigenvalue.

We first let $1 \leq i_1 < \cdots < i_k \leq n$. Because $\tilde{\boldsymbol{W}} \in \mathbb{R}^{d\times n}$ and $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$, we have the following:

$$\begin{aligned}\prod_{t=1}^k \sigma_{i_t}(\tilde{\boldsymbol{W}}\boldsymbol{Q}) &= \prod_{t=1}^k \sqrt{\lambda_{i_t}(\tilde{\boldsymbol{W}}\boldsymbol{Q}\boldsymbol{Q}^\top\tilde{\boldsymbol{W}}^\top)} \\ &= \sqrt{\prod_{t=1}^k \lambda_{i_t}(\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}\boldsymbol{Q}\boldsymbol{Q}^\top)}\end{aligned} \tag{57}$$

by applying Lemma 4 to the above equation, we have that

$$\begin{aligned}\sqrt{\prod_{t=1}^{k}\lambda_{i_t}(\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}\boldsymbol{Q}\boldsymbol{Q}^\top)} &\geq \sqrt{\prod_{t=1}^{k}\lambda_{i_t}(\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}})\lambda_{n-t+1}(\boldsymbol{Q}\boldsymbol{Q}^\top)} \\ &= \prod_{t=1}^{k}\sigma_{i_t}(\tilde{\boldsymbol{W}})\sigma_{n-t+1}(\boldsymbol{Q})\end{aligned} \tag{58}$$

$$\begin{aligned}\sqrt{\prod_{t=1}^{k}\lambda_{i_t}(\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}\boldsymbol{Q}\boldsymbol{Q}^\top)} &\leq \sqrt{\prod_{t=1}^{k}\lambda_{i_t}(\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}})\lambda_{t}(\boldsymbol{Q}\boldsymbol{Q}^\top)} \\ &= \prod_{t=1}^{k}\sigma_{i_t}(\tilde{\boldsymbol{W}})\sigma_{t}(\boldsymbol{Q})\end{aligned} \tag{59}$$

Therefore, we have that

$$\prod_{t=1}^{k}\sigma_{i_t}(\tilde{\boldsymbol{W}}\boldsymbol{Q}) \geq \prod_{t=1}^{k}\sigma_{i_t}(\tilde{\boldsymbol{W}})\sigma_{n-t+1}(\boldsymbol{Q}) \tag{60}$$

$$\prod_{t=1}^{k}\sigma_{i_t}(\tilde{\boldsymbol{W}}\boldsymbol{Q}) \leq \prod_{t=1}^{k}\sigma_{i_t}(\tilde{\boldsymbol{W}})\sigma_{t}(\boldsymbol{Q}) \tag{61}$$

Suppose we have $k=1$ and $i_1=n$, then Eq. (60) gives

$$\sigma_n(\tilde{\boldsymbol{W}}\boldsymbol{Q}) \geq \sigma_n(\tilde{\boldsymbol{W}})\sigma_n(\boldsymbol{Q}) \tag{62}$$

Then suppose we have $k=1$ and $i_1=1$, then Eq. (61) gives

$$\sigma_1(\tilde{\boldsymbol{W}}\boldsymbol{Q}) \leq \sigma_1(\tilde{\boldsymbol{W}})\sigma_1(\boldsymbol{Q}) \tag{63}$$

Combining the above results with Eq. (52) and Eq. (54), we have that

$$\begin{aligned}\lim_{n\to\infty}\sigma_{\max}(\boldsymbol{W}) = \lim_{n\to\infty}\sigma_{\max}(\tilde{\boldsymbol{W}}\cdot\boldsymbol{Q}) \leq \lim_{n\to\infty}\big(\sigma_{\max}(\tilde{\boldsymbol{W}})\cdot\sigma_{\max}(\boldsymbol{Q})\big) = (\sqrt{d}+\sqrt{\lambda d})\cdot\max_i\frac{1}{\|\tilde{\boldsymbol{v}}_i\|_1} \\ \lim_{n\to\infty}\sigma_{\min}(\boldsymbol{W}) = \lim_{n\to\infty}\sigma_{\min}(\tilde{\boldsymbol{W}}\cdot\boldsymbol{Q}) \geq \lim_{n\to\infty}\big(\sigma_{\min}(\tilde{\boldsymbol{W}})\cdot\sigma_{\min}(\boldsymbol{Q})\big) = (\sqrt{d}-\sqrt{\lambda d})\cdot\min_i\frac{1}{\|\tilde{\boldsymbol{v}}_i\|_1}\end{aligned} \tag{64}$$

which concludes the proof. □

J Hyperspherical Uniformity from Zero-mean Gaussian Distributions

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

Theorem 7. *The normalized vector of Gaussian variables is uniformly distributed on the sphere. Formally, let* $x_1, x_2, \cdots, x_n \sim \mathcal{N}(0,1)$ *and be independent. Then the vector*

$$\boldsymbol{x} = \left[\frac{x_1}{z}, \frac{x_2}{z}, \cdots, \frac{x_n}{z}\right] \tag{65}$$

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

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

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{66}$$

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

$$f(x) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\langle x, x\rangle}. \tag{67}$$

Then we introduce the following lemma (Lemma 5) about the orthogonal-invariance of the normal distribution.

Lemma 5. *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 (*$\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$*). Then* $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{x}$ *also has the distribution of* $\mathcal{N}(0,1)$*.*

Proof. For any measurable set $A \subset \mathbb{R}^n$, we have that

$$\begin{aligned} P(Y \in A) &= P(X \in U^\top A) \\ &= \int_{U^\top A} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\langle x, x\rangle} \\ &= \int_{A} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\langle Ux, Ux\rangle} \\ &= \int_{A} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\langle x, x\rangle} \end{aligned} \tag{68}$$

because of orthogonality of U. Therefore the lemma holds. □

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 $\mathbb{N}(0,1)$ and then projecting it onto the hypersphere $\mathbb{S}^{n-1}$ produces random vectors $U = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ that are uniformly distributed on the hypersphere. Therefore the theorem holds. □

K Orthogonality vs. Orthonormality

In the paper, we sometimes use the term "orthogonality" and "orthonormality" interchangeably, since we are mostly considering the points lying in $\mathbb{S}^d$. Hyperspherical uniformity only concerns with the angles among points (*e.g.*, neurons), because all the magnitude are normalized to one before entering any hyperspherical uniformity objective. Therefore strictly speaking, orthogonality is a more appropriate comparison to hyperspherical uniformity.

L Experimental Details

Layer	CNN-9 for CIFAR-100	ResNet-32 for CIFAR-100	ResNet-18 for ImageNet-2012
Conv0.x	N/A	[3×3, 64]	[7×7, 64], Stride 2 3×3, Max Pooling, Stride 2
Conv1.x	[3×3, 64]×3 2×2 Max Pooling, S2	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 5$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$
Conv2.x	[3×3, 128]×3 2×2 Max Pooling, S2	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 5$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$
Conv3.x	[3×3, 256]×3 2×2 Max Pooling, S2	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 5$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$
Conv4.x	N/A	N/A	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$
Final	256-Dim Fully Connected	Average Pooling	

Table 7: Our CNN and ResNet architectures with different convolutional layers. Conv0.x, Conv1.x, Conv2.x, Conv3.x and Conv4.x denote convolution units that may contain multiple convolutional layers, and residual units are shown in double-column brackets. Conv1.x, Conv2.x and Conv3.x usually operate on different size feature maps. These networks are essentially similar to [73] and [31], but with different number of filters in each layer. The downsampling is performed by convolutions with a stride of 2. E.g., [3×3, 64]×4 denotes 4 cascaded convolution layers with 64 filters of size 3×3, and S2 denotes stride 2.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense $\rightarrow M_g \times M_g \times 512$
4×4, stride=2 deconv. BN 256 ReLU
4×4, stride=2 deconv. BN 128 ReLU
4×4, stride=2 deconv. BN 64 ReLU
3×3, stride=1 conv. 3 Tanh

(a) Generator ($M_g = 4$ for CIFAR10).

RGB image $x \in \mathbb{R}^{M \times M \times 3}$
3×3, stride=1 conv 64 lReLU 4×4, stride=2 conv 64 lReLU
3×3, stride=1 conv 128 lReLU 4×4, stride=2 conv 128 lReLU
3×3, stride=1 conv 256 lReLU 4×4, stride=2 conv 256 lReLU
3×3, stride=1 conv. 512 lReLU
dense $\rightarrow$ 1

(b) Discriminator ($M = 32$ CIFAR10).

Table 8: GAN architecture for image generation on CIFAR-10. The architectures mostly follow [57].

General. For MHE, we use half-space MHE with $s = 2$. For the unrolling of MHP and MHC, we use one-step gradient descent to approximate the inner optimization. For MHC, we use the relaxed formulation with $\gamma = 5$. For MGD, we use Gaussian kernel with $\epsilon = 1$. Typically, we search the best weighting hyperparameter for all the regularizations from 10^{-8} to 10^{7} (with 10 as the step size).

Multilayer perceptron. We conduct hand-written digit recognition task on MNIST with a three-layer multilayer perceptron following this repository[1] . The size of each digit image is 28×28, which is 784 dimensions after flattened. Both hidden layers have 256 output dimensions, *i.e.*, 256 neurons. The output layer will output 10 logits for classification. Finally, we use a cross-entropy loss with softmax function. We use the momentum SGD optimizer with learning rate 0.01, momentum 0.9 and batch size 100. The training stops at 100 epochs.

Convolutional neural networks. The network architectures used in the main paper are specified in Table 7. For all experiments, we use the momentum SGD optimizer with momentum 0.9. For CIFAR-100, we set the mini-batch size as 128. The learning rate starts at 0.1, and is divided by 10 when the performance is saturated. For ImageNet-2012, we use the mini-batch size 128 and the training starts with learning rate 0.1. The learning rate is divided by 10 when the performance is saturated, and the training is terminated at 700k iterations. The structure of ResNet-18 mostly follows [31]. Note that, for all the methods in our experiments, we always use the best possible hyperparameters for the corresponding regularization (via cross-validation) to make sure that the comparison is fair. The baseline has exactly the same training settings as the others. Standard ℓ_2 weight decay $(5e-4)$ is applied by default to all the methods.

Graph networks. We implement the all the hyperspherical uniformity regularizations for GCN in the official repository[2]. All the hyperparameter settings exactly follow this official repository to ensure a fair comparison.

Point cloud networks. To simplify the comparison and remove all the bells and whistles, we use a vanilla PointNet (without T-Net) as our backbone network. We apply OPT to train the MLPs in PointNet. We follow the same experimental settings as [62] and evaluate on the ModelNet-40 dataset [82]. We exactly follow the

[1] https://github.com/hwalsuklee/tensorflow-mnist-MLP-batch_normalization-weight_initializers
[2] https://github.com/tkipf/gcn

same setting in the original paper [62] and the official repositories[3]. Specifically, we use the hyperspherical uniformity regularizations to regularize all the 1×1 convolution layers and the fully connected layer (except the final classifier). For the experiments, we set the point number as 1024 and mini-batch size as 32. We use the Adam optimizer with initial learning rate 0.001. The learning rate will decay by 0.7 every 200k iterations, and the training is terminated at 250 epochs.

Generative adversarial networks. The architecture we use for the GAN experiments is shown in Table 8. For fair comparison, all the hyperparameter settings exactly follow [57]. We use leaky ReLU (LReLU) in the newtork and set the slopes of LReLU functions to 0.1.

L.1 Experimental details for Fig. 2

For the experiment in Fig. 2, we use 200 3-dimensional neurons. The momentum SGD optimizer (with momentum 0.9) is used to optimize these hyperspherical uniformity objectives. The learning rate starts at 0.01 and is divided by 10 at 5k iterations. The optimization stops at 8k iterations. In Fig. 2(a), the y-axis denotes the value of hyperspherical energy. In Fig. 2(b), the y-axis denotes the value of separation distance. We did not visualize MHP, since the true objective value of MHP is difficult (also time-consuming) to compute. For MHE, we use half-space MHE with $s = 2$. For MHC, we use one-step gradient descent to approximate the inner optimization and also adopt the relaxed formulation with $\gamma = 5$. For MGD, we use Gaussian kernel with $\epsilon = 1$.

L.2 Experimental details for Fig. 3

For the visualization experiment in Fig. 3, we optimize 100 3-dimensional neurons on the unit sphere. The training hyperparameters are the same as Section L.1.

[3] https://github.com/charlesq34/pointnet