Stochastic Embeddings: A Probabilistic and Geometric Analysis of Out-of-Distribution Behavior

Anthony Nguyen[1,2] **Emanuel Aldea**[2] **Sylvie Le Hégarat-Mascle**[2] **Renaud Lustrat**[1]

[1]Thales Land and Air Systems, BU ARC, Limours, France
[2]SATIE Laboratory UMR 8029, Paris-Saclay University, CNRS, Gif-sur-Yvette, France

Abstract

Deep neural networks perform well in many applications but often fail when exposed to out-of-distribution (OoD) inputs. We identify a geometric phenomenon in the embedding space: in-distribution (ID) data show higher variance than OoD data under stochastic perturbations. Using high-dimensional geometry and statistics, we explain this behavior and demonstrate its application in improving OoD detection. Unlike traditional post-hoc methods, our approach integrates uncertainty-aware tools, such as Bayesian approximations, directly into the detection process. Then, we show how considering the unit hypersphere enhances the separation of ID and OoD samples. Our mathematically sound method achieves competitive performance while remaining simple.

1 INTRODUCTION

Machine learning models are widely used in fields such as healthcare, autonomous systems, and natural language processing. Deploying them in real-world applications poses challenges often overlooked during development. For instance, a key challenge is detecting when models are uncertain or encounter unfamiliar inputs. Despite their strong performance on clean datasets, deep neural networks often overestimate confidence on unknown or degraded inputs [Guo et al., 2017]. This raises reliability concerns, particularly in sensitive applications where models encounter unexpected data or distributions not seen during training.

Uncertainty quantification has become crucial for evaluating model predictions. Early methods, inspired by Bayesian statistics [Robert, 2005], led to approaches like Deep Ensemble [Lakshminarayanan et al., 2017] and Bayesian Neural Network approximations. These methods were later adapted for OoD detection [Malinin and Gales, 2018, Charpentier et al., 2021]. OoD detection focuses on identifying inputs that do not fit the statistical features of the training data. Such inputs may correspond to novel or anomalous situations where the model's predictions could be unreliable. Simple methods based on Softmax confidence scores Hendrycks and Gimpel [2016] have shown limitations, as Deep Neural Networks (DNNs) often give overconfident predictions, even for artificial OoD inputs [Hein et al., 2019].

Recently, several deterministic methods have been proposed to quantify uncertainty [Van Amersfoort et al., 2020, Mukhoti et al., 2022, Nguyen et al., 2024], often using distances or local density in the embedding space. These methods focus on its geometry, which becomes complex in high-dimensional settings [Nalisnick et al., 2019]. Recent theoretical advances have examined and exploited the geometry [Papyan et al., 2020, Pearce et al., 2021, Ammar et al., 2024] and analytical properties [Tian et al., 2021, Haas et al., 2023] induced by the Cross-Entropy (CE) loss function to enhance OoD detection. These insights show how some structural properties of the basic CE Loss can provide a fruitful way to enhance separation between ID and OoD inputs. Concurrently, competitive post-hoc methods on pre-trained networks [Djurisic et al., 2023, Sun and Li, 2022, Sun et al., 2021] have shown simplicity and strong OoD detection performance. Some of these methods exploit embedding geometric properties [Lee et al., 2018, Sun et al., 2022] and are often more efficient than methods requiring additional training [Zhang et al., 2023]. Despite these advances in probabilistic and deterministic methods, detecting OoD samples accurately while ensuring interpretability and robustness remains challenging [Yang et al., 2023, Jaeger et al., 2022]. Our contributions are as follows:

1. **Exploration of the variance behavior when injecting stochasticity into the embedding space**: we investigate a counter-intuitive observation arising from the application of Monte Carlo (MC) Dropout within the embedding space, instead of the more commonly studied logit space. One would expect OoD samples to exhibit greater variance across multiple stochastic forward passes, reflecting higher uncertainty

compared to ID data during inference. However, our empirical results show the opposite: ID samples consistently exhibit higher variance than OoD samples under MC Dropout.

2. **Mathematical explanation**: Using high-dimensional probability theory and differential geometry, we explain this variance behavior through the geometric properties of the hypersphere and isotropic random vectors. We show how this insight improves OoD detection.

3. **A simple and effective algorithm**: We present an algorithm that delivers excellent performance on standard OoD benchmarks. It is easy to implement, robust in high dimensions, and supported by solid mathematical foundations.

2 RELATED WORK

2.1 STOCHASTIC UNCERTAINTY QUANTIFICATION

Uncertainty quantification evaluates how confident a model is in its predictions by accounting for both the limitations of the model itself and the variability present in the data [Hullermeier and Waegeman, 2021]. A common Bayesian-inspired method for estimating uncertainty is Monte Carlo (MC) Dropout. MC Dropout approximates Bayesian inference by applying Dropout during inference and sampling multiple stochastic forward passes through the network:

$$\mathbb{P}(y|\mathbf{x},\mathcal{D}_\gamma) = \int \mathbb{P}(y|\mathbf{x},\theta)\mathbb{P}(\theta|\mathcal{D}_\gamma)d\theta, \qquad (1)$$

where $\mathcal{D}_\gamma$ denotes the training dataset, $\mathbf{x}$ the input and y the label. Formally, for classification with K classes for an input $\mathbf{x}$, M stochastic forward passes during the inference phase yields a set of outputs $\{P(y|\mathbf{x},\theta_1),\ldots,P(y|\mathbf{x},\theta_M)\}$ and performs the approximation of Eq.(1) by the following empirical mean:

$$\forall y \in [\![1,K]\!], \frac{1}{M}\sum_{i=1}^{M} P(y|\mathbf{x},\theta_i) \simeq \mathbb{P}(y|\mathbf{x},\mathcal{D}_\gamma). \qquad (2)$$

Despite its success, Dropout may be limited in capturing sufficient diversity in predictions because it focuses only on deactivating neuron outputs. To address this, DropConnect [Wan et al., 2013] provides a more effective mechanism for inducing diversity by injecting fine-grained noise directly into the weight matrices, and producing more nuanced perturbations in the network's embeddings.

Let $U \in \mathcal{M}_{K\times D}(\mathbb{R})$ be the weight matrix of a fully connected layer, so that for an input $\mathbf{x} \in \mathbb{R}^D$ the output is $\mathbf{z} = U\mathbf{x}$. DropConnect introduces stochasticity directly into the weight matrix as follows:

1. Generate a binary mask matrix B $\in \{0,1\}^{K\times D}$, where each entry is sampled independently as $B_{ij} \sim \text{Bernoulli}(p)$ and $p = 1-q$, with q being the probability of $B_{ij} = 1$.

2. Compute the effective weight matrix via the Hadamard product: $\widetilde{U} = U \odot B$. For an input $\mathbf{x} \in \mathbb{R}^D$, the layer's output becomes $\tilde{\mathbf{z}} = \widetilde{U}\mathbf{x} = (U \odot B)\mathbf{x}$.

2.2 DETERMINISTIC UNCERTAINTY QUANTIFICATION

2.2.1 Local density, distance and curse of dimensionality

To detect OoD samples, a simple idea is to measure the uncertainty of the samples and to classify as OoD those whose uncertainty exceeds a certain threshold. For instance, prior deterministic methods such as those proposed in Van Amersfoort et al. [2020], Mukhoti et al. [2022] define uncertainty in terms of the distance or local density of an input relative to training samples in the embedding space. An input $\mathbf{x}$ is assumed to belong to the training distribution if its embedding $\mathbf{z} = h_\theta(\mathbf{x})$ lies near a class-specific centroid $\boldsymbol{\mu}_c$. Thus, the uncertainty score is defined as:

$$\text{Uncertainty}(\mathbf{x}) \propto \min_c \|\mathbf{z} - \boldsymbol{\mu}_c\|^2. \qquad (3)$$

An input is classified as OoD if its minimum distance to any class centroid exceeds a threshold. This approach links large distances in the embedding space to higher uncertainty. However, it can be limited by the curse of dimensionality [Vershynin, 2018].

In high-dimensional embedding spaces, the discriminative power of distances to class centroids or local density can diminish, so simple thresholds become less effective at separating ID and OoD samples. This phenomenon is a well-documented manifestation of the curse of dimensionality, where increasing feature dimensions can erode the meaningfulness of distance and density metrics, even though the neural network produces well-separated clusters in the embedding space [Olteanu et al., 2023]. Recent works start taking this into account explicitly. For instance, SIREN [Du et al., 2022] projects the embeddings into a smaller-dimensional space and then normalize them on the hypersphere to fit a von Mises–Fisher distribution. Nguyen et al. [2024] likewise use a projection to reduce dimensionality when describing the embedding's geometry.

2.2.2 Analytical methods

Recent analytical OoD detection approaches instead intervene directly in the network's internal representations or constraining activation patterns to more reliably handle OoD inputs. Recent works from Sun et al. [2021], Azizmalayeri et al. [2024] modify the embedding activations through clipping above a high percentile threshold based on ID statistics to directly suppress the excessive signals often produced by OoD inputs. Djurisic et al. [2023] similarly, truncate activations beyond a certain percentile and proportionally scale the rest to diminish the impact of hypersensitive neuron. Alternatively, works from Haas et al. [2023], Wei et al. [2022] scale embedding and pre-softmax logits respectively during training. More precisely, Haas et al. [2023] scale embedding vectors so that their norms more faithfully reflect each input's difficulty. LogitNorm method [Wei et al., 2022] by contrast, rescales pre-softmax logits, observing that even when most training examples are already classified correctly, the softmax cross-entropy loss keeps driving logit norm large, leading to overconfidence.

2.3 GEOMETRY OF THE EMBEDDING

Several studies from Pearce et al. [2021], Tian et al. [2021] have examined the geometric and analytical properties of the embedding space induced by the CE loss to improve OoD detection. CE loss promotes class separation by creating well-defined geometric structures within the embedding space, where samples from the same class are tightly clustered and different classes are well-separated. This phenomenon, known as **Neural Collapse** (NC), described by Papyan et al. [2020] and illustrated in Fig. 1, occurs in the final stages of training. NC describes the convergence of class embeddings to well-separated class means, or centroids, while the within-class variance decreases. Specifically, the embeddings of samples within the same class collapse to their respective class means, and the class means themselves align symmetrically in a spatially equi-distributed/repartitioned way that maximizes inter-class separation. Additionally, the class vectors align with the embeddings, so that each representation points toward its corresponding class prototype.

3 PRELIMINARIES

This section introduces the notation and background used throughout the paper. We define key symbols and provide the mathematical framework underlying our study.

3.1 HYPOTHESES AND BACKGROUND

Let the training set and the testing set be denoted as $\mathcal{D}_{\text{Train}} = \{(\mathbf{x}_i, y_i), i \in [\![1, N_{\text{Train}}]\!]\}$ and $\mathcal{D}_{\text{Test}} = \{(\mathbf{x}_i, y_i), i \in [\![1, N_{\text{Test}}]\!]\}$ respectively. Here $\mathbf{x}_i \in \mathbb{R}^p$ represents an image and $y_i \in [\![1, K]\!]$ its associated label where K stands for the total number of classes. We assume that both datasets are independently and identically distributed (i.i.d.) according to their respective joint distributions $\mathbb{P}_{\text{Train}} := \mathbb{P}_{\text{Train}}(\mathbf{x}, y)$ and $\mathbb{P}_{\text{Test}} := \mathbb{P}_{\text{Test}}(\mathbf{x}, y)$.

Out-of-Distribution (OoD): We assume that the training and test sets follow a common distribution denoted by $\mathbb{P}_{\text{ID}}$ (ID data). We introduce another test set of OoD samples $\{(\mathbf{x}_i, \upsilon_i), i \in [\![1, N_{\text{OoD}}]\!]\}$ which are drawn i.i.d. from an unknown distribution denoted by $\mathbb{P}_{\text{OoD}}$, distinct from $\mathbb{P}_{\text{ID}}$.

In the context of image classification, the embedding of an input image $\mathbf{x}$ is defined by $\mathbf{z} = h_\theta(\mathbf{x}) \in \mathbb{R}^D$ where D denotes the dimension of the embedding space. Inputs, embedding related vectors, and class vectors are written in bold, $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^D x_i^2}$ denotes the L^2 norm of the vector $\mathbf{x} \in \mathbb{R}^D$ and $S^{D-1} := \{\mathbf{x} \in \mathbb{R}^D \mid \|\mathbf{x}\| = 1\}$ denotes the unit hypersphere in $\mathbb{R}^D$.

The model is divided into two components: the feature extractor denoted by h_θ and a final linear layer g_θ acting as a classifier. Thus, the DNN's output can be written as $f_\theta(\mathbf{x}) = g_\theta \circ h_\theta(\mathbf{x})$. Since $g_\theta : \mathbb{R}^D \to \mathbb{R}^K$ is a linear operator, it can be expressed as a weight matrix $W_\theta \in \mathcal{M}_{K,D}(\mathbb{R})$.

To introduce stochasticity during the inference phase, consider a fixed input $\mathbf{x}$. Let $\{\mathbf{z}^{(m)}, m \in [\![1, M]\!]\}$ represent the collection of embeddings obtained from M stochastic forward passes through the network. Each embedding is defined as $\mathbf{z}^{(m)} := h_\theta(\mathbf{x}; \sigma^{(m)}) \in \mathbb{R}^D$, where $\sigma^{(m)}$ denotes the stochastic perturbation applied during the m-th forward pass.

If $\mathbf{x}_{\text{ID}} \sim \mathbb{P}_{\text{ID}}(\mathbf{x})$ (resp. $\mathbf{x}_{\text{OoD}} \sim \mathbb{P}_{\text{OoD}}(\mathbf{x})$, we denote by $Z_{\text{ID}} \in \mathcal{M}_{D,M}(\mathbb{R})$ (resp. Z_{OoD}) the matrix whose columns are the vectors $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}$, omitting the index M to simplify the notation.

If DropConnect is applied to produce M vectors, the matrix with these vectors as its columns is denoted by Z_{DC} or A_{DC} depending on the context.

To quantify the dispersion of these families of vectors, for any matrix $Z \in \mathcal{M}_{D,M}(\mathbb{R})$, we define the non-biased estimator:

$$\text{Var}(Z) := \text{Tr}\left(\frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{z}^{(i)} - \boldsymbol{\mu})(\mathbf{z}^{(i)} - \boldsymbol{\mu})^T\right), \tag{4}$$

where Tr is the Trace operator of a matrix, i.e., the sum of its diagonals entries and

$$\boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{z}^{(i)}. \tag{5}$$

3.2 BEHAVIOR OF THE EMBEDDING UNDER CROSS ENTROPY OPTIMIZATION

3.2.1 Geometrical behavior

Consider a deterministic DNN taking an image $\mathbf{x} \in \mathbb{R}^p$ and generating an embedding $\mathbf{z} = h_\theta(\mathbf{x}) \in \mathbb{R}^D$. This embedding is then passed through the classification layer, producing a logit vector $\boldsymbol{\ell} := (\ell_1, ..., \ell_K) \in \mathbb{R}^K$. The logits are then normalized using the Softmax function:

$$\forall k \in [\![1, K]\!], \mathbb{P}(y = k|\mathbf{x}) = \frac{\exp \ell_k}{\sum_{i=1}^K \exp \ell_i} \in [0, 1]. \tag{6}$$

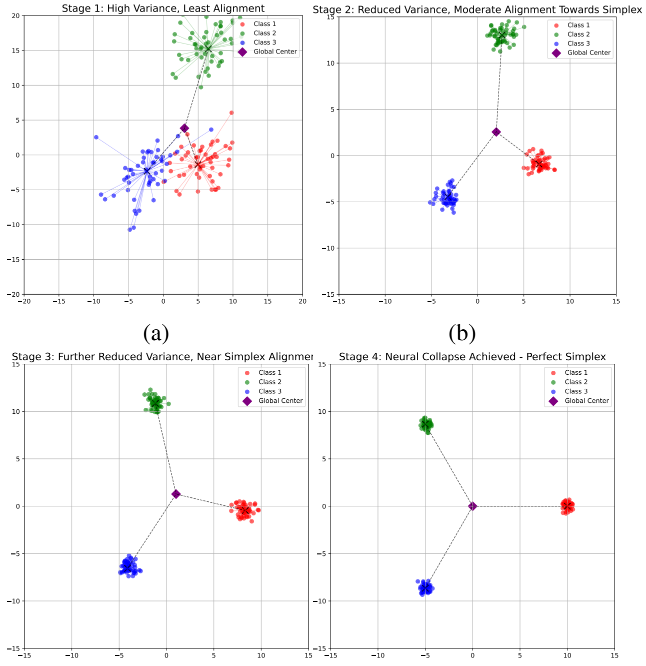
For a given sample $(\mathbf{x}, y) \in \mathbb{R}^p \times [\![1, K]\!]$, the CE loss used for backpropagation is defined as

$$\mathcal{L}_{CE}(\mathbf{x}, y) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\text{Train}}}[-\log \mathbb{P}(y|\mathbf{x})]. \quad (7)$$

The NC phenomenon studied by Papyan et al. [2020] and shown in Fig. 1, describes how class embeddings converge to well-separated centroids in the later training stages.

In fact, NC is not something unusual or particular, but rather a natural phenomenon due to its mathematical basis and empirical consistency in supervised learning. Lu and Steinerberger [2022] provide theoretical justification, showing that NC arises as an optimal configuration under cross-entropy minimization. Additionally, Graf et al. [2021] extend this understanding by observing that supervised contrastive loss also leads to similar geometric configurations, indicating that this structure emerges naturally and consistently across different optimization paradigms in deep learning.

Regarding OoD behavior under Neural Collapse, Pearce et al. [2021] demonstrated that when the data exhibits low aleatoric uncertainty and the feature extractor is sufficiently deep, the simplex configuration depicted in Fig. 1d is both achievable and optimal for OoD detection using MSP [Hendrycks and Gimpel, 2016], as OoD embedding samples tend to cluster near the origin and around the decision boundaries. This finding is further supported by Ammar et al. [2024].



(a)

(b)

(c)

(d)

Figure 1: Illustration of Neural Collapse with the progressive emergence of a simplex configuration from 1a to 1d.

3.2.2 Analytical behavior

In Eq. (6), $\forall k \in [\![1, K]\!], \ell_k$ can be expressed using cosine similarity and the classifier's weight vectors. Let $\mathbf{w}_k$ denote the k-th columns vectors of W_θ and $\phi_k = \arccos\left(\frac{\mathbf{w}_k^T \mathbf{z}}{\|\mathbf{w}_k\|\|\mathbf{z}\|}\right)$, then, the logit is given by:

$$\ell_k = \mathbf{w}_k^T \mathbf{z} = \|\mathbf{w}_k\|\|\mathbf{z}\| \cos(\phi_k). \quad (8)$$

Substituting this into the Softmax probability expression, we have:

$$\mathbb{P}(y = k|\mathbf{x}) = \frac{\exp(\|\mathbf{w}_k\|\|\mathbf{z}\| \cos(\phi_k))}{\sum_{i=1}^K \exp(\|\mathbf{w}_i\|\|\mathbf{z}\| \cos(\phi_i))}. \quad (9)$$

Tian et al. [2021] hypothesizes that the confidence assigned to an input's most likely class is strongly influenced by the norm of its feature representation. However, because the norm is unconstrained, it may become less sensitive to the difficulty of the input.

Algorithm 1 Normalization of Features

```
function FORWARD($x$)
    $\mathbf{z} \leftarrow h_\theta(\mathbf{x})$
    featurenorm $\leftarrow \|\mathbf{z}\|$
    $\mathbf{z} \leftarrow \frac{\mathbf{z}}{\|\mathbf{z}\|}$
    $y \leftarrow g_\theta(\mathbf{z})$
    return $y$, featurenorm
end function
```

To address this, Haas et al. [2023] propose applying L^2 normalization to the embedding features $\mathbf{z} = h_\theta(\mathbf{x})$ transforming them into $\frac{\mathbf{z}}{\|\mathbf{z}\|}$ before computing the logits. This step decouples the feature magnitudes from equinormality constraints. By normalizing the embeddings only during training, the method preserves variability in feature norms, allowing them to better capture input-specific difficulty.

Importantly, doing so ensures that the feature norms of OoD samples are much lower than ID's, making them an effective indicator for OoD detection. This work is the foundation of our method. To the reader's convenience, the normalization is presented in Algorithm 1.

4 STOCHASTIC EMBEDDING DYNAMICS

As discussed in Section 2, adding stochasticity to the final layer alone does not directly provide an effective solution for OoD detection. To explore its potential benefits, we first examine a DNN trained using the CE loss with Dropout on the embedding.

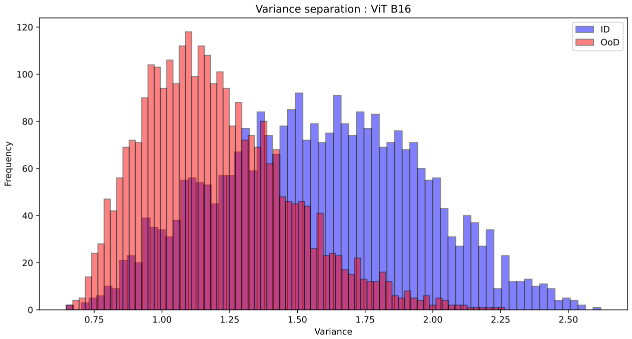


Figure 2: Histograms of $\text{Var}(Z_{\text{ID}})$ (in blue) and $\text{Var}(Z_{\text{OoD}})$ (in red), derived by applying Dropout during the inference phase to generate Z_{ID} for ID inputs (ImageNet) and Z_{OoD} for OoD inputs (Textures).

A notable observation, shown in Fig. 2, is that applying MC Dropout to the penultimate layer during inference consistently results in $\text{Var}(Z_{\text{OoD}}) < \text{Var}(Z_{\text{ID}})$. This result may seem counterintuitive, as OoD inputs are usually expected to exhibit higher variance.

4.1 ASSUMPTIONS

To understand why $\text{Var}(Z_{\text{OoD}}) < \text{Var}(Z_{\text{ID}})$ is observed, we analyze the embedding space under specific assumptions.

Considering our trained DNN, the first step is to introduce the set of OoD samples detected by MSP during inference:

$$\begin{aligned} D_{\text{OoD}}^{\text{MSP}} = &\Big\{(\mathbf{x}, y) \mid \mathbf{z} = h_\theta(\mathbf{x}),\ \|\mathbf{z}\| \leq \tau\Big\} \\ &\cup \Big\{(\mathbf{x}, y) \mid \mathbf{z} = h_\theta(\mathbf{x}),\ \|\mathbf{z}\| > \tau\Big\}, \end{aligned} \tag{10}$$

for some $\tau \in \mathbb{R}_+^*$. We focus on OoD samples with low feature norms:

$$D_{\text{OoD}}^{\text{MSP}}(\tau) := \Big\{(\mathbf{x}, y) \mid \mathbf{z} = h_\theta(\mathbf{x}),\ \|\mathbf{z}\| \leq \tau\Big\}.$$

Let $\varepsilon \in [\varepsilon_{\text{min}}, 1]$, where

$$\varepsilon_{\text{min}} = 1 - \max_{\mathbf{x}, j} \frac{\mathbf{w}_j^\top h_\theta(\mathbf{x})}{\|\mathbf{w}_j\| \|h_\theta(\mathbf{x})\|}.$$

Finally, let us partition $D_{\text{OoD}}^{\text{MSP}}(\tau)$ into

$$\begin{aligned} D_{\text{OoD}}^{\text{MSP}}(\tau) = &\Big\{(\mathbf{x}, y) \mid \exists j \in [\![1, K]\!] : \cos(\phi_j) \geq 1 - \varepsilon\Big\} \\ &\cup \Big\{(\mathbf{x}, y) \mid \forall j \in [\![1, K]\!] : \cos(\phi_j) < 1 - \varepsilon\Big\}, \end{aligned}$$

For the sake of theoretical study, if we suppose that:

- The DNN f_θ is trained using the regular CE loss,
- Then NC occurs along with the configurations described by Pearce et al. [2021],

then we can safely assume that the set defined in the following Lemma 4.1 is non-empty.

Lemma 4.1 (See Appendix A). *Let $\mathbf{x}_{\text{ID}} \sim \mathbb{P}_{\text{ID}}(\mathbf{x})$ be an ID sample that is correctly classified by MSP as ID such as $\exists k \in [\![1, K]\!], \cos(\phi_k) = 1$, and define $\mathbf{z}_{\text{ID}} = h_\theta(\mathbf{x}_{\text{ID}})$. Let $(\mathbf{x}_{\text{OoD}}, \upsilon)$ be an OoD sample such that $\mathbf{z}_{\text{OoD}} = h_\theta(\mathbf{x}_{\text{OoD}})$ and*

$$(\mathbf{x}_{\text{OoD}}, \upsilon) \in \Big\{(\mathbf{x}, y) \mid \exists j \in [\![1, K]\!] :\ \cos(\phi_j) \geq 1 - \varepsilon\Big\}, \tag{11}$$

Then, for a suitably chosen τ, we have

$$\|\mathbf{z}_{\text{OoD}}\| \leq \|\mathbf{z}_{\text{ID}}\|. \tag{12}$$

Proof. We refer the reader to Appendix C. □

Now, the next step is to incorporate geometric and probabilistic concepts to model the Dropout effect when applied to the embedding during inference. as MC Dropout applied to an embedding $\mathbf{z} = \|\mathbf{z}\|\boldsymbol{\varphi}$ perturbs both its norm and its direction.

We first analyze the scenario where only the directional component is affected, as modeled by Theorem 4.2.

4.2 SPHERICAL CAP GEOMETRY AND ITS ROLE IN EMBEDDING DISPERSION

Theorem 4.2 (see Appendix A). *Let $\mathbf{z} \in \mathbb{R}^D$ be an embedding vector, and write*

$$\mathbf{z} = \|\mathbf{z}\|\boldsymbol{\varphi}, \tag{13}$$

for some fixed unit vector $\boldsymbol{\varphi} \in S^{D-1}$. Let $\Phi \in [0, \pi]$ be given, and define the spherical cap

$$C_\Phi(\boldsymbol{\varphi}) \ = \ \Big\{\boldsymbol{\alpha} \in S^{D-1} : \boldsymbol{\alpha}^\top \boldsymbol{\varphi} \geq \cos\Phi\Big\}. \tag{14}$$

The concentration parameter $c = \mathbb{E}\big[\boldsymbol{\alpha}^\top \boldsymbol{\varphi}\big]$ quantifies how tightly the perturbed directions are distributed around $\boldsymbol{\varphi}$.

For $M \in \mathbb{N}^*$ we suppose $\{\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \ldots\ \boldsymbol{\alpha}^{(M)}\}$ is a sequence of i.i.d. random vectors on $C_\Phi(\boldsymbol{\varphi})$.

We define the matrix:

$$Z_M := \|\mathbf{z}\| A_M := \|\mathbf{z}\| \left(\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \ldots, \boldsymbol{\alpha}^{(M)}\right), \tag{15}$$

and we denote $\text{Var}(Z_M)$ its variance estimator. Then the following properties hold:

1. ***Finite Expectation:** For all $M \geq 1$,*

$$\mathbb{E}\Big[\text{Var}(Z_M)\Big] = \|\mathbf{z}\|^2 \left(1 - c^2\right). \tag{16}$$

2. ***Almost-Sure Asymptotics:*** *As $M \to \infty$,*

$$\mathrm{Var}(Z_M) \xrightarrow{a.s.} \|\mathbf{z}\|^2 \left(1 - c^2\right). \tag{17}$$

Proof. We refer the reader to Appendix C. □

To interpret the role of c^2, note that when $M \to +\infty$ and $c^2 = 0$, the perturbed directions are uniformly distributed over the hypersphere. Conversely, if $c^2 = 1$, the perturbations are fully concentrated around $\boldsymbol{\varphi}$.

For simplicity, we assumed i.i.d random vectors to be uniformly distributed over $C_\Phi(\boldsymbol{\varphi})$, i.e., locally uniform over the hypersphere. Of course, this assumption may be extended by adopting any relevant statistical model with a well-defined variance supported on the hypersphere.

Now, we have all the necessary components to explain the observation in Fig. 2 .

4.3 HOW FEATURE NORMS AMPLIFY VARIANCE IN ID DATA

Let $\mathbf{x}_{\text{ID}} \sim \mathbb{P}_{\text{Test}}(\mathbf{x})$ and $\mathbf{x}_{\text{OoD}} \sim \mathbb{P}_{\text{OoD}}(\mathbf{x})$ be samples whose original embeddings denoted by $\mathbf{z}_{\text{ID, orig}}$ and $\mathbf{z}_{\text{OoD, orig}}$ satisfy the conditions of Lemma 4.1, such that $\|\mathbf{z}_{\text{OoD, orig}}\| \leq \|\mathbf{z}_{\text{ID, orig}}\|$.

Under the model of Theorem 4.2 where only the direction is perturbed and the original norm is fixed, we have:

$$\mathbb{E}[\mathrm{Var}(Z_{\text{ID}})] = \|\mathbf{z}_{\text{ID, orig}}\|^2(1 - c_{\text{ID}}^2), \tag{18}$$

$$\mathbb{E}[\mathrm{Var}(Z_{\text{OoD}})] = \|\mathbf{z}_{\text{OoD, orig}}\|^2(1 - c_{\text{OoD}}^2). \tag{19}$$

Assuming comparable angular dispersion i.e., $c_{\text{ID}}^2 = c_{\text{OoD}}^2$ due to standard training not differentiating this aspect for the considered samples, Lemma 4.1 implies

$$\mathbb{E}[\mathrm{Var}(Z_{\text{OoD}})] \leq \mathbb{E}[\mathrm{Var}(Z_{\text{ID}})] \tag{20}$$

We now extend this to the general case where MC Dropout perturbs both the norm and direction. Let $\mathbf{z}'$ be an embedding after application of the Dropout mask, such that $\mathbf{z}' = s\boldsymbol{\alpha}\|\mathbf{z}_{\text{orig}}\|$, where $s \in [0, 1]$ is a stochastic norm scaling factor and $\boldsymbol{\alpha}$ is the stochastic unit direction. We assume the following:

1. **Independent Norm Scaling:** The random variable S (for s_m) is independent of the original norm and has the same distribution for ID and OoD samples. Let $\kappa = \mathbb{E}[S^2]$, where $0 < \kappa \leq 1$. Thus, the average squared post-Dropout norm is $\mathbb{E}_{\text{masks}}[\|\mathbf{z}'\|^2] = \mathbb{E}[S^2\|\mathbf{z}_{\text{orig}}\|^2] = \kappa\|\mathbf{z}_{\text{orig}}\|^2$.
2. **Decoupled uniform Perturbations:** While s and $\boldsymbol{\alpha}$ both arise from the same Dropout mask, we approximate that the variance structure from Theorem 4.2 can be applied by replacing the fixed $\|\mathbf{z}\|^2$ with $\mathbb{E}_{\text{masks}}[\|\mathbf{z}'\|^2]$.

Under these conditions, the average post-Dropout norms are:

$$\mathbb{E}_{\text{masks}}[\|\mathbf{z}'_{\text{ID}}\|^2] = \kappa\|\mathbf{z}_{\text{ID, orig}}\|^2 \tag{21}$$

$$\mathbb{E}_{\text{masks}}[\|\mathbf{z}'_{\text{OoD}}\|^2] = \kappa\|\mathbf{z}_{\text{OoD, orig}}\|^2 \tag{22}$$

Since $\|\mathbf{z}_{\text{OoD, orig}}\|^2 \leq \|\mathbf{z}_{\text{ID, orig}}\|^2$ and $\kappa > 0$, it follows that $\mathbb{E}_{\text{masks}}[\|\mathbf{z}'_{\text{OoD}}\|^2] \leq \mathbb{E}_{\text{masks}}[\|\mathbf{z}'_{\text{ID}}\|^2]$. If the angular concentration parameters c_{ID}^2 and c_{OoD}^2 remain comparable, i.e., $c_{\text{ID}}^2 = c_{\text{OoD}}^2$, then:

$$\mathbb{E}[\mathrm{Var}(Z_{\text{ID}})] = \kappa\|\mathbf{z}_{\text{ID, orig}}\|^2(1 - c_{\text{ID}}^2) \tag{23}$$

$$\mathbb{E}[\mathrm{Var}(Z_{\text{OoD}})] = \kappa\|\mathbf{z}_{\text{OoD, orig}}\|^2(1 - c_{\text{OoD}}^2) \tag{24}$$

This leads to $\mathbb{E}[\mathrm{Var}(Z_{\text{OoD}})] \leq \mathbb{E}[\mathrm{Var}(Z_{\text{ID}})]$.

While our model, particularly the decoupling approximation, simplifies the complex interaction of norm and directional perturbations from Dropout, it provides a rationale for the observed variance difference. The inherent dual impact of Dropout nonetheless complicates general ID/OoD separation, as suggested by phenomena in Fig. 2 and Fig. 6.

In conclusion, our analysis suggest that while MC Dropout introduces stochasticity into the embeddings, it does so in an uncontrolled way by perturbing both the norm and the direction simultaneously. This mixing of effects leads to the observed higher variance for ID samples primarily due to their overall larger norms for the class of samples considered.

5 NORM AND ANGULAR DECOUPLING

It might be tempting to optimize the variance difference i.e., to force $\mathrm{Var}(Z_{\text{OoD}}) \ll \mathrm{Var}(Z_{\text{ID}})$ as a means to distinguish between ID and OoD data. **Instead**, our strategy pursues an alternative approach that does not rely on enhancing such variance differences but rather consists of decoupling norm from the angular component of the embedding vector:

1. We impose some constraints on the angular concentration, a parameter that remained unconstrained in the standard Dropout setup described in Sec. 4. To achieve this, we first add a fully-connected DropConnect layer after the embedding [Wan et al., 2013], then apply normalization, thereby leveraging the Central Limit Theorem and concentration of measure phenomena.
2. A fortunate byproduct of this design is that we naturally integrate the L2-normalization strategy from Haas et al. [2023], as detailed in Algorithm 1.

5.1 TRAINING PHASE

During the training step, for each input $\mathbf{x}$ passed through the feature extractor $h_\theta(\mathbf{x})$, we get an embedding vector $\mathbf{z}$. To introduce random rotation, a fully-connected stochastic

Algorithm 2 Training Phase

```
**Input (Training):** Train input $\mathbf{x}$, Feature extractor $h_\theta(\cdot)$, DropConnect function $\mathrm{DC}(\cdot)$, Classifier $g_\theta(\cdot)$
**for** each batch of data $\mathbf{x}$ **do**
    $\mathbf{z} \leftarrow h_\theta(\mathbf{x})$
    $r \leftarrow \|\mathbf{z}\|$
    $\boldsymbol{\alpha} \leftarrow \mathbf{z}/r$
    $\boldsymbol{\alpha}_{DC} \leftarrow \mathrm{DC}(\boldsymbol{\alpha})$
    $\boldsymbol{\ell} \leftarrow g_\theta(\frac{\boldsymbol{\alpha}_{DC}}{\|\boldsymbol{\alpha}_{DC}\|})$
    Compute loss $\mathcal{L}$ using $\boldsymbol{\ell}$ and labels
    Back-propagate to update network weights $\theta$
**end for**
```

linear layer DC(.) utilizing DropConnect and matching the dimensionality of $\mathbf{z}$ is added after the embedding layer.

The normalized embedding is passed through the DropConnect function $\mathrm{DC}(\cdot)$, producing a stochastically perturbed vector $\boldsymbol{\alpha}_{DC}$. Since the output is not guaranteed to be a unit vector, we normalize it again. The DC layer, combined with normalization, stretches, distorts, and projects the vector $\boldsymbol{\alpha}$ onto the hypersphere.

Using DropConnect means that the fully connected layer $DC : \mathbb{R}^D \to \mathbb{R}^D$ outputs each component $\boldsymbol{\alpha}_i$ of $\boldsymbol{\alpha}_{DC}$ as a sum of many independent and uniformly bounded contributions, each multiplied by a Bernoulli random variable. Consequently, by the Central Limit Theorem through the Lindeberg's condition [Lindeberg, 1922], each component of $\boldsymbol{\alpha}_{DC}$ asymptotically satisfies:

$$\forall i \in [\![1, D]\!], \sqrt{D}\,(\boldsymbol{\alpha}_{DC})_i \xrightarrow{d} \mathcal{N}(\delta_i, \sigma_i^2) \qquad (25)$$

as $D \to +\infty$ and $\xrightarrow{d}$ denotes convergence in distribution. Since the components are independent, the entire random vector $\boldsymbol{\alpha}_{DC}$ is asymptotically Gaussian. Consequently, $\boldsymbol{\alpha}_{DC}$ behaves as a Gaussian random vector with diagonal covariance matrix, and the normalized version $\frac{\boldsymbol{\alpha}_{DC}}{\|\boldsymbol{\alpha}_{DC}\|}$ is distributed over a spherical cap. To simplify the presentation, we kept assuming that the normalized vector is uniformly distributed over a spherical cap. A more precise study of this statistical model with its true distribution is provided in Appendix D.

Training in this way creates meaningful angular differences between ID and OoD. Indeed, the exposition of ID data to angular perturbation during the training refines the network, making the model invariant to the specific angular perturbations introduced by DropConnect and effectively confining ID inputs within a smaller spherical cap (i.e., $c_{\mathrm{ID}}^2 \simeq 1$) as illustrated in Appendix, Fig. 5 and observed in Fig. 3.

5.2 INFERENCE PHASE

At inference time, we keep the DropConnect stochasticity active by applying the DC layer immediately after the embedding layer, followed by normalization, using the same DropConnect rate as during training. This results in M different perturbations of the angle $\boldsymbol{\alpha}_{DC}^{(i)}, i \in [\![1, M]\!]$, all associated with the same norm $\|\mathbf{z}\|$ which is held constant across perturbations.

Algorithm 3 Inference Phase and OoD Detection

```
**Input:** Test input $\mathbf{x}$, Feature extractor $h_\theta(\cdot)$, DropConnect function $\mathrm{DC}(\cdot)$, Classifier $g_\theta(\cdot)$, Number of forward passes $M$
$\mathbf{z} \leftarrow h_\theta(\mathbf{x})$
$r \leftarrow \|\mathbf{z}\|$
$\boldsymbol{\alpha} \leftarrow \mathbf{z}/r$
**for** $m = 1$ **to** $M$ **do**
    $\boldsymbol{\alpha}_{DC}^{(m)} \leftarrow \mathrm{DC}(\boldsymbol{\alpha})$
    $\boldsymbol{\alpha}_{DC}^{(m)} \leftarrow \frac{\boldsymbol{\alpha}_{DC}^{(m)}}{\|\boldsymbol{\alpha}_{DC}^{(m)}\|}$
**end for**
In a validation set, compute $\bar{r} = \frac{1}{IQR \times N}\sum_{i=1}^{N} r_i$, $r_i$ is the norm of the $i$-th element and $IQR$ is the Interquartile Range.
Define the OoD score as $S_{DC}(\mathbf{x}) := \mathrm{Var}(A_{DC}) + \lambda\frac{\bar{r}-r}{r}$.
**Output:** OoD Score $S_{DC}(\mathbf{x})$
```

After completing the M passes, the variance of the matrix $A_{DC} = (\boldsymbol{\alpha}_{DC}^{(1)}, ..., \boldsymbol{\alpha}_{DC}^{(M)})$ is calculated. If $\mathbf{z} = h_\theta(\mathbf{x})$ and $r = \|\mathbf{z}\|$, we define the score as

$$S_{DC}(\mathbf{x}) := \mathrm{Var}(A_{DC}) + \lambda\frac{\bar{r} - r}{r}, \qquad (26)$$

where $\bar{r}$ is computed as the mean norm divided by the interquartile range (IQR) of the norms on a validation set with $IQR = Q(0.75) - Q(0.25)$, and $Q(p)$ is the p-th quantile.

Dividing by the IQR makes the norm score robust to outliers and provides a consistent scaling factor that reflects both the central tendency and variability of the validation set.

The hyperparameter λ can be chosen, for instance, as the 90th percentile (or another appropriate quantile) of the score distribution computed on ID data from the validation set.

Note that using only $\mathrm{Var}(Z)$ to separate ID from OoD led to poor and unstable performance likely due to a mismatch in the optimization objective. Indeed, $\mathrm{Var}(Z_{\mathrm{ID}}) = r^2\,\mathrm{Var}(A_{\mathrm{ID}})$ and while r^2 increases for ID data, $\mathrm{Var}(A_{\mathrm{ID}})$ decreases. The opposite occurs for $\mathrm{Var}(Z_{\mathrm{OoD}})$.

6 EXPERIMENTS

We applied a high DropConnect rate on the linear DropConnect layer, with empirical results showing **that** $p \in [0.8, 0.9]$ **yields optimal performance**.

During the inference phase, we used $M = 50$ forward passes to compute $\mathrm{Var}(Z)$. While this number may initially appear low given the size of the embedding space, working on the unit hypersphere allows us to benefit of the blessing of dimensionality namely, the concentration of measure in high-dimensional spaces, which ensure that even a moderate number of passes provides a reliable estimation of the variance, as shown in Appendix B.2 and D.

6.1 EFFECT OF DROPCONNECT ON THE HYPERSPHERE

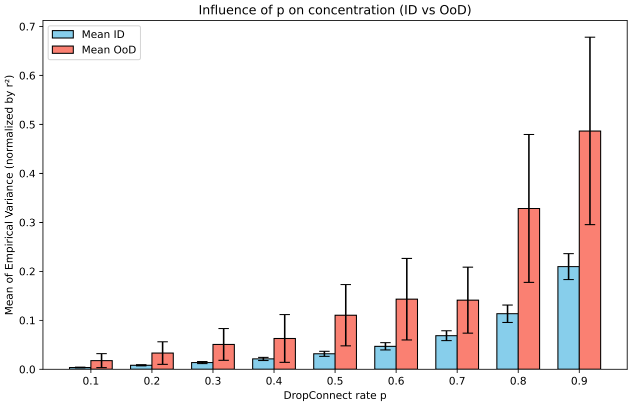


Figure 3: Evolution of $1 - c^2$ as $p \to 1$.

To derive Fig. 3, we trained the same model using various DropConnect rates $p \in \{0.1, \ldots, 0.9\}$. For each fixed p and for every test input $\mathbf{x} \sim \mathbb{P}_{\text{Test}}(\mathbf{x})$, we performed $M = 50$ forward passes without rescaling by the norm, thereby obtaining a $1 - c^2$ value per input. We then computed the mean and standard deviation of these values across the entire test and OoD datasets. The error bars indicate the standard deviation. Finally, we plotted these estimates as a function of p. As $p \to 1$, the angular component concentration becomes clearly separated and statistically significant (e.g., for $p = 0.8$ and $p = 0.9$) between ID and OoD data. We also observe that all test ID inputs exhibit tightly concentrated c^2_{ID} values, consistent with the concentration of measure in high dimensions. In contrast, the c^2_{OoD} values for OoD inputs show significantly more dispersion, indicating a **weaker** concentration phenomena, suggesting that the angular response of the model to OoD data is more variable and less predictable.

6.2 OOD DETECTION

We built our OoD benchmark around embedding-based, post-hoc detectors chosen for their ease of integration and proven effectiveness. These techniques are prevalent in the OoD community and serve as a solid foundation for our tests, with several recognized as state-of-the-art on large-scale architectures. To evaluate the effectiveness of these methods, we report three metrics: the area under the ROC curve (AUROC), the area under the precision–recall curve (AUPRC), and the false positive rate at 95% true positive rate (FPR95). As shown in Tables 1, 2 and 3, our composite score consistently ranks among the top three across diverse benchmarks, demonstrating robustness and generality across datasets. It is worth noting that DeepKNN [Sun et al., 2022] often tops these benchmarks largely because it normalizes every embedding before performing the k-nearest-neighbours. That norm–based separation amplifies the gap between ID and OoD points, probably giving DeepKNN an edge. Indeed, Sun et al. [2021], Azizmalayeri et al. [2024] observed that OoD activations tend to be sparser than ID activations, so normalizing these vectors may make them appear even more sparse on the unit sphere. The increased sparsity of OoD activations, once passed through the DC layer, may be responsible for the observed amplification of angular variance and may explain why our method naturally amplifies the angular-variance term for OoD inputs.

Table 1: OOD detection on CIFAR-100 (ID) → SVHN and CIFAR-10 (OOD) using ResNet-18.

	SVHN		
Method	**AUROC**	**AUPRC**	**FPR@95TPR**
MSP	84.69	86.67	57.28
MaxLogit	83.57	86.96	76.96
ReAct	83.12	83.33	57.01
Energy Score	86.51	57.79	98.40
ASH B	84.09	83.01	58.00
ASH P	84.22	81.88	56.11
ASH S	88.01	83.10	54.99
DeepKNN	93.61	**94.15**	52.43
DDU	80.63	57.09	94.57
Norm (Feature)	86.95	91.73	50.69
ViM	92.81	92.66	**49.67**
Mahalanobis	82.21	83.03	91.01
Naive Sampling	72.33	76.27	60.28
LogitNorm	82.27	59.66	86.45
Ours	**94.23**	93.85	65.23

	CIFAR-10		
Method	**AUROC**	**AUPRC**	**FPR@95TPR**
MSP	76.75	81.21	73.87
MaxLogit	73.79	80.14	90.94
Energy Score	75.51	80.44	97.80
ReAct	76.44	82.11	75.44
ASH B	72.20	81.01	74.17
ASH P	**77.90**	82.84	76.04
ASH S	71.93	82.11	75.14
DeepKNN	77.66	**83.15**	71.91
Energy Score	75.51	80.44	97.80
DDU	74.74	80.68	97.96
Norm (Feature)	76.37	80.76	**70.01**
ViM	76.01	82.19	81.33
Mahalanobis	73.92	82.03	85.01
Naive Sampling	69.24	77.44	76.30
LogitNorm	74.78	79.49	73.03
Ours	77.27	82.92	70.46

7 CONCLUSION

We present an exploratory study and a mathematically grounded method for enhancing OoD detection. This work focuses on exploratory analysis and modeling to explain the geometric and probabilistic phenomena observed in embedding spaces. Our exploration revealed that when applying MC Dropout to the embedding layer, ID samples tended to exhibit higher variance than OoD samples primarily due to their larger feature norms. This observation highlighted a critical limitation: MC Dropout affects both norm and angle in an uncontrolled manner, which obscures the true uncertainty signal needed to differentiate between ID and OoD data. By establishing a link between uncertainty and concentration of measure, our OoD score integrates controlled angular variance using DropConnect and a norm-based component, leveraging both directional and magnitude information in the embeddings. We hope this connection will offer useful insights and stimulate further interest.

Table 2: OOD detection on CIFAR-10 (ID) → SVHN and CIFAR-100 (OoD) using ResNet-18.

SVHN

Method	**AUROC**	**AUPRC**	**FPR@95TPR**
MSP	87.17	92.59	38.02
MaxLogit	90.70	95.41	45.84
Energy Score	90.94	52.46	99.78
ReAct	87.57	92.22	44.02
ASH B	79.44	84.01	63.01
ASH P	83.99	89.12	54.56
ASH S	82.01	92.11	49.03
DeepKNN	95.19	97.26	**9.83**
DDU	84.09	56.70	87.85
Norm (Feature)	94.89	97.68	24.22
ViM	95.17	**98.68**	21.05
Mahalanobis	88.45	67.34	79.12
Naive Sampling	81.11	83.55	88.11
LogitNorm	93.05	70.66	80.45
Ours	**95.37**	98.52	18.60

CIFAR-100

Method	**AUROC**	**AUPRC**	**FPR@95TPR**
MSP	80.62	77.54	72.62
MaxLogit	75.90	76.65	78.34
Energy Score	75.93	64.18	99.47
ReAct	81.77	77.19	72.12
ASH B	74.11	71.01	72.21
ASH P	85.99	86.12	64.56
ASH S	81.01	82.11	66.99
DeepKNN	**88.51**	**86.30**	40.33
DDU	83.55	66.49	98.74
Norm (Feature)	87.98	86.09	**40.01**
ViM	87.52	85.68	50.05
Mahalanobis	84.79	71.44	91.01
Naive Sampling	77.10	68.11	91.43
LogitNorm	82.78	63.49	80.03
Ours	88.01	86.27	47.77

Table 3: OOD detection on ImageNet (ID) vs three OOD sets (ResNet-50).

NINCO

Method	**AUROC**	**AUPRC**	**FPR@95TPR**
MSP	83.20	58.87	67.79
MaxLogit	86.67	64.52	52.85
Energy Score	81.85	61.01	99.82
ReAct	81.61	48.19	73.11
ASH-P	78.54	55.78	66.54
ASH-B	91.04	74.04	55.67
ASH-S	88.56	**79.11**	44.11
Norm(Feature)	87.49	69.37	40.87
DeepKNN	**93.80**	77.12	**14.06**
ViM	92.14	73.56	25.21
Mahalanobis	85.23	71.83	49.36
DDU	83.12	67.93	41.22
Naive Sampling	79.45	59.74	55.09
LogitNorm	92.22	71.59	22.30
Ours	93.61	76.19	21.03

Textures

Method	**AUROC**	**AUPRC**	**FPR@95TPR**
MSP	69.32	60.59	85.30
MaxLogit	75.81	64.92	83.14
Energy Score	27.11	52.27	99.75
ReAct	74.12	64.91	90.12
ASH-P	83.42	70.32	85.46
ASH-B	65.24	59.73	99.53
ASH-S	79.93	66.98	77.22
Norm(Feature)	80.79	65.72	76.14
DeepKNN	**85.06**	73.97	62.58
ViM	84.55	72.09	**60.12**
Mahalanobis	77.02	58.13	93.84
DDU	79.33	71.36	81.22
Naive Sampling	71.74	58.22	65.27
LogitNorm	84.88	69.46	62.00
Ours	84.97	**74.10**	68.11

Places365

Method	**AUROC**	**AUPRC**	**FPR@95TPR**
MSP	73.58	85.40	79.86
MaxLogit	75.68	86.47	78.67
Energy Score	66.30	82.23	98.46
ReAct	75.11	84.89	91.12
ASH-P	85.08	91.01	79.02
ASH-B	77.10	81.71	85.10
ASH-S	79.56	88.96	81.03
Norm(Feature)	82.46	89.56	76.45
DeepKNN	84.41	89.28	60.33
ViM	**85.97**	88.56	**41.39**
Mahalanobis	75.18	84.20	80.70
DDU	73.44	80.36	92.39
Naive Sampling	69.40	79.95	89.32
LogitNorm	84.82	89.28	38.26
Ours	85.10	**91.23**	64.10

References

Mouin Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. *ICLR*, 2024.

Mohammad Azizmalayeri, Ameen Abu-Hanna, and Giovanni Cinà. Mitigating overconfidence in out-of-distribution detection by capturing extreme activations. *arXiv preprint arXiv:2405.12658*, 2024.

Djalil Chafaï and Joseph Lehec. Logarithmic sobolev inequalities essentials. *Accessed on*, page 4, 2024.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian uncertainty for exponential family distributions. *arXiv preprint arXiv:2105.04471*, 2021.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *ICLR*, 2023.

Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. *Advances in Neural Information Processing Systems*, 35:20434–20449, 2022.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Jarrod Haas, William Yolland, and Bernhard Rabus. Exploring simple, high quality out-of-distribution detection with l2 normalization. 2023.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 41–50, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Eyke Hullermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3): 457–506, 2021.

Paul F Jaeger, Carsten T Lüth, Lukas Klein, and Till J Bungert. A call to reflect on evaluation practices for failure detection in image classification. *arXiv preprint arXiv:2211.15259*, 2022.

Bum Jun Kim, Hyeyeon Choi, Hyeonah Jang, Donggeon Lee, and Sang Woo Kim. How to use dropout correctly on residual networks with batch normalization. In *Uncertainty in Artificial Intelligence*, pages 1058–1067. PMLR, 2023.

Alexander V Kolesnikov and Emanuel Milman. Riemannian metrics on convex sets with applications to poincaré and log-sobolev inequalities. *Calculus of Variations and Partial Differential Equations*, 55(4):77, 2016.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Jarl Waldemar Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, pages 211–225, 1922. Volume contains 328 pages. Online archive available.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

J Mukhoti, A Kirsch, J van Amersfoort, PHS Torr, and Y Gal. Deep deterministic uncertainty: A simple baseline. arxiv, 2022.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Hai-Vy Nguyen, Fabrice Gamboa, Reda Chhaibi, Sixin Zhang, Serge Gratton, and Thierry Giaccone. Combining statistical depth and fermat distance for uncertainty quantification. *Advances in Neural Information Processing Systems*, 2024.

Madalina Olteanu, Fabrice Rossi, and Florian Yger. Meta-survey on outlier and anomaly detection. *Neurocomputing*, 555:126634, 2023.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Tim Pearce, Alexandra Brintrup, and Jun Zhu. Understanding softmax confidence and uncertainty. *arXiv preprint arXiv:2106.04972*, 2021.

Christian Robert. *Le choix bayésien: Principes et pratique*. Springer Science & Business Media, 2005.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

Junjiao Tian, Dylan Yung, Yen-Chang Hsu, and Zsolt Kira. A geometric perspective towards neural calibration via sensitivity decomposition. *Advances in Neural Information Processing Systems*, 34:26358–26369, 2021.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022.

William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-ood: Deciphering modern out-of-distribution detection algorithms. *arXiv preprint arXiv:2310.01755*, 2023.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Li Hai. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Stochastic Embeddings : A Probabilistic and Geometric Analysis of Out-of-Distribution Behavior

Anthony Nguyen[1,2] **Emanuel Aldea**[2] **Sylvie Le Hégarat-Mascle**[2] **Renaud Lustrat**[1]

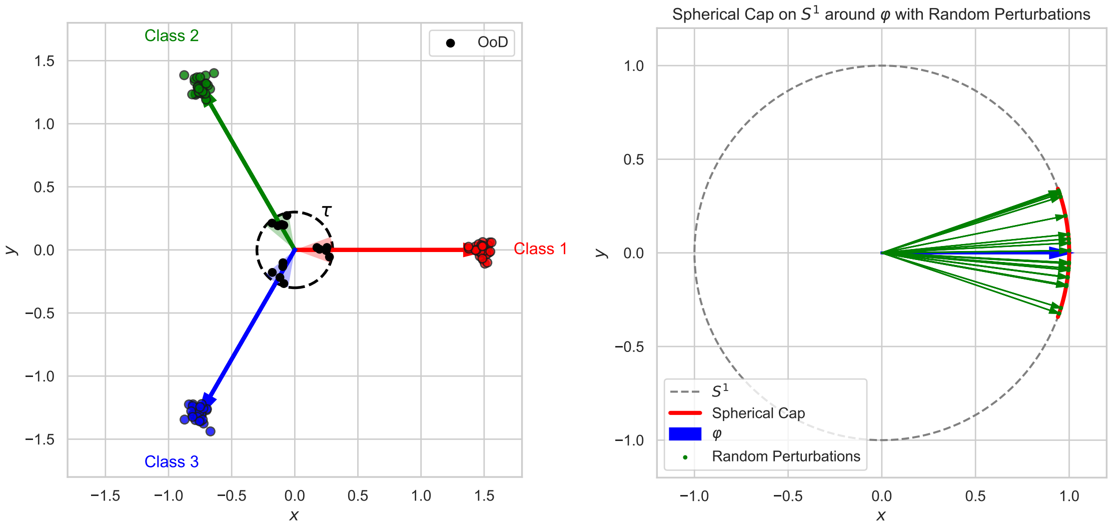
[1]Thales Land and Air Systems, BU ARC, Limours, France
[2]SATIE Laboratory UMR 8029, Paris-Saclay University, CNRS, Gif-sur-Yvette, France

A VISUALIZATION

Fig. 4 shows respectively the illustrations of the considered setting in Lemma 4.1 and the mathematical spherical cap defined in Theorem 4.2. In particular, Fig. 4(a) illustrates the embedding's simplex configuration where the OoD embedding lies around the origin with high cosine similarity (and where ID data are clustered along the class vector), while Fig. 4(b) depicts the spherical cap centered around the original embedding direction vector φ, within which sampling is performed uniformly.

In Fig. 5 left (resp. right) blue arrow represents the initial direction φ of the vector $\mathbf{z} = \|\mathbf{z}\|\varphi \in \mathbb{P}_{\text{Test}}(\mathbf{x})$ (resp. $\mathbb{P}_{\text{OoD}}(\mathbf{x})$) . During inference, the green (resp. purple) arrows represents the M perturbed vectors, induced by M stochastic forward passes. Our method computes the variance on all these green vectors. Φ represent the spherical cap limits. Same in the right picture. As the illustration shows, when stochastically perturbed during inference, ID embeddings exhibit greater stability under stochastic perturbation than OoD embeddings, i.e., $c^2_{\text{OoD}} \le c^2_{\text{ID}}$.

Fig. 6 indicate a high correlation between the norm separation of the embeddings (see Fig. 6a) and the variance separation (see Fig. 6b) under MC Dropout, as studied in Sec. 4.



(a) Illustration for Lemma 4.1.

(b) Illustration of the 2D spherical cap for Theorem 4.2.

Figure 4: (a) Simplex configuration where OoD embedding lies around the origin with high cosine similarity and ID data are clustered along the class vector. (b) Spherical cap defined in Theorem 4.2.

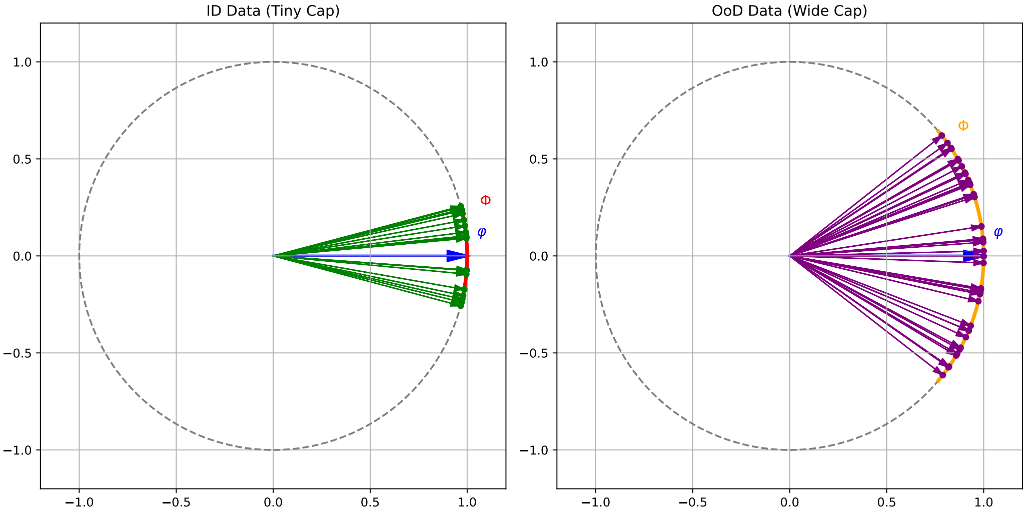
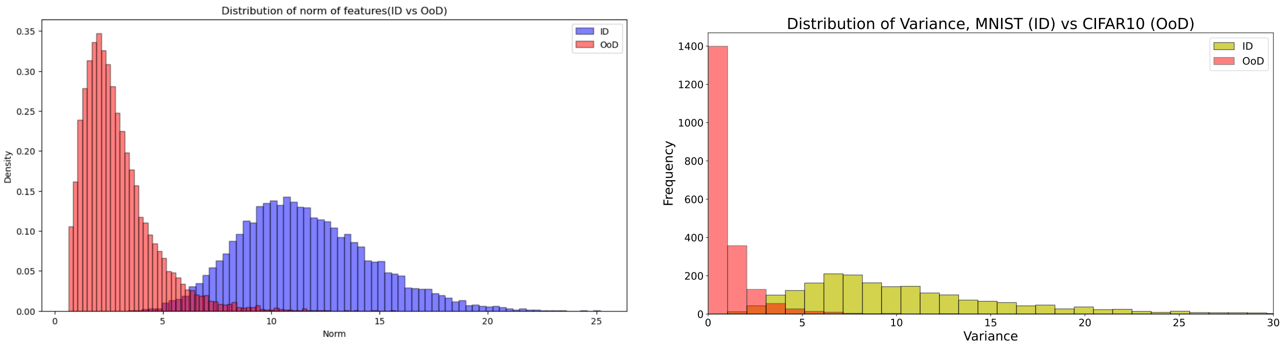


Figure 5: Illustration of the post-training behavior: ID data exhibit more concentration as the DNN remains invariant to stochastic-induced perturbations.



(a) Norm ID (MNIST) vs Norm OoD(CIFAR10).

(b) Variance ID (MNIST) vs. Variance OoD (CIFAR10).

Figure 6: Separation of ID/OoD embeddings on toy data. Great norm separation leads to great variance separation.

B ADDITIONAL EXPERIMENTS AND DETAILS

B.1 STOCHASTICITY ON INTERMEDIATE LAYERS

We introduced stochasticity into intermediate layers using Dropout (following Kim et al. [2023]). We evaluated the performance by comparing AUROC scores, with the injected stochasticity propagating through the feature extractor's output during inference. We believe that the results shown in Table 4 are not surprising, as intermediate layers typically capture lower-level features that are less discriminative for distinguishing between ID and OoD data.

Table 4: **ResNet18 : CIFAR10 (Id) vs CIFAR100(OoD)**.

Modified layer	**AUROC (%)**
layer1	52.3
layer2	53.4
layer3	63.0
layer4	71.1
Embedding	**80.1**

B.2 DROPCONNECT PARAMETER SENSITIVITY AND COMPUTATIONAL COST

In all of our experiments we use the same DropConnect rate p during both training and inference, to ensure consistency of the trainable parameters in the stochastic layer. While our main paper does not include a detailed empirical study of how p affects in–distribution (ID) accuracy, we now present such results for completeness.

Table 5: CIFAR–10 and CIFAR–100 ID accuracy (%) as a function of DropConnect rate p.

DropConnect rate p	**CIFAR–10 ID Acc.**	**CIFAR–100 ID Acc.**	**ImageNet ID Acc.**
0.1	91.83	71.21	76.11
0.2	91.53	72.44	75.99
0.3	91.76	71.98	75.52
0.4	90.96	71.90	75.40
0.5	91.28	71.12	74.83
0.6	91.26	72.35	75.76
0.7	91.32	70.56	75.33
0.8	91.09	71.04	75.11
0.9	91.03	70.12	75.43

ID accuracy vs. DropConnect rate In the main paper we select $p = 0.9$ to enhance OOD separation on the hypersphere (see Fig. 3). noting that slightly lower values of p can yield marginal gains in ID accuracy but at the cost of degraded OOD performance.

Training time overhead Higher DropConnect rates incur slower convergence during training. We measure the relative increase in wall-clock training time (to reach the same validation loss) as seen in Table 6:

Table 6: Relative training time increase (%) vs. DropConnect rate p.

Drop prob. p	**Training time ↑ (%)**
0.1	0.0%
0.2	6.5%
0.3	12.7%
0.4	6.5%
0.5	12.3%
0.6	19.1%
0.7	26.5%
0.8	28.4%
0.9	36.2%

Inference cost vs. number of passes Unlike standard MC techniques, our multiple stochastic passes can be started from the embedding layer, reducing cost. Table 7 shows average batch times (128 images) for a full forward-backward pass vs. our optimized partial-forward strategy:

Table 7: Average inference time per batch for M passes (ResNet-50, 128 images).

M	**Full pass (s)**	**Optimized (s)**	**Speedup**
1	0.0244	0.0241	1.01×
5	0.1203	0.0401	3.00×
10	0.2404	0.0600	4.01×
15	0.3672	0.0804	4.57×
20	0.4829	0.0994	4.86×
30	0.7279	0.1386	5.25×
40	0.9715	0.1781	5.45×
50	1.2144	0.2176	5.58×

Variance concentration vs. number of passes Finally, we report in Table 8 how quickly the empirical variance of an ID sample converges to the reference value as M increases (averaged over 500 samples):

Table 8: Convergence of average empirical variance vs. M (500 samples).

M	**Avg. variance on CIFAR10**	**Avg. variance on ImageNet**
10	0.2323	0.3372
15	0.2308	0.3303
20	0.2349	0.3217
25	0.2321	0.3144
30	0.2201	0.3113
35	0.2118	0.3098
40	0.2116	0.3107
45	0.2113	0.3104
50	0.2121	0.3110

B.3 VON-MISES FISHER CONCENTRATION ON THE UNIT HYPERSPHERE

Alternatively, to further validate that during inference ID data exhibit higher concentration on the unit hypersphere, we characterized this concentration on the unit hypersphere using a Von Mises-Fisher distribution. We trained our model using DropConnect rate $p = 0.5$ then applied MC DropConnect during inference as described in Algorithm 3. The density of the Von Mises–Fisher distribution is defined as follows:

$$f_D(\mathbf{x}) := C_D(\kappa)\exp(\kappa\psi\mathbf{x}),\ \forall\mathbf{x} \in S^{D-1}, \tag{27}$$

where $\|\psi\| = 1, \kappa \geq 0$, and the normalization constant $C_D(\kappa)$ is equal to :

$$C_D(\kappa) := \frac{\kappa^{D/2-1}}{(2\pi)^{D/2} I_{D/2-1}(\kappa)}, \tag{28}$$

where I_v denotes the modified Bessel function. The greater the value of κ, the higher the distribution is concentrated around ψ.

We observe in Fig. 7 that ID data is clustering more tightly than OoD data on the unit hypersphere, though this is not optimal due to the insufficient DropConnect rate. Consequently, the concentration parameter κ may serve as a valuable metric for further analysis.

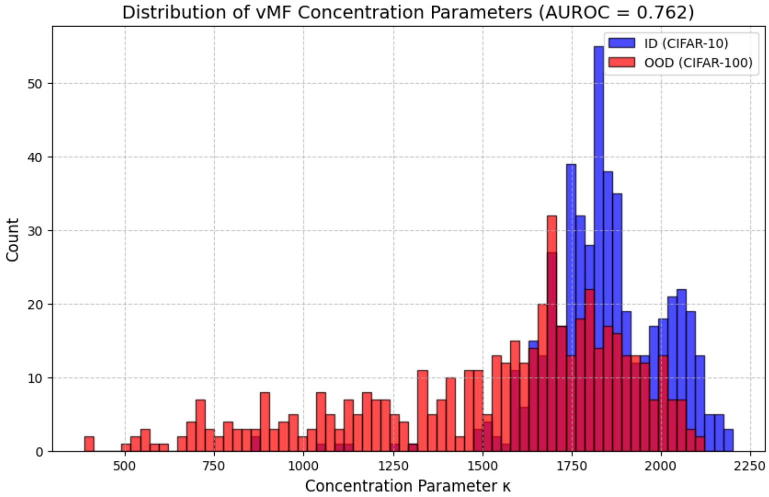


Figure 7: Concentration κ is used as OoD Score.

B.4 FINE-TUNING DETAILS

- For toy datasets (MNIST) we used a multi-layer perceptron with a DropConnect layer in its third layer. The architecture is 784-256-256(DC)-128-10 with ReLU activations. We trained for 40 epochs using SGD with 0.01 learning rate and 1×10^{-3} Weight Decay.
- CIFAR10/CIFAR100: we fine-tuned a vanilla ResNet18 model (pretrained in PyTorch) with the first convolutional layer modified to use a 3×3 kernel. We trained for 200 epoch using SGD, 128 batch-size,and momentum of 0.9, with DropConnect rate of 0.9 on the DC layer which is has the same structure and being fully connected to the penultimate layer. We used an initial learning rate of 0.1 with a cosine annealing scheduler and we applied standard data augmentation techniques : cropping, horizontal flipping.
- ImageNet: we fine-tuned a vanilla ResNet50 model (pretrained in PyTorch). We trained for 150 epoch using SGD, 128 batch-size,and momentum of 0.9, with DropConnect rate of 0.9 on the DC layer which is has the same structure and being fully connected to the penultimate layer. We used an initial learning rate of 0.1 with a cosine annealing scheduler and we applied standard data augmentation techniques : cropping, horizontal flipping.
- For the ViT visualization, since it was pretrained on ImageNet, we had to fine-tune all layers to achieve high accuracy. We trained for 30 epochs with a batch size of 64, SGD, used a weight decay of 5×10^{-4}, set the momentum to 0.9, and applied Dropout with a probability of 0.5 on the penultimate layer. We used an initial learning rate of 0.01 and we applied standard data augmentation techniques : cropping, horizontal flipping.

C PROOF OF THEORETICAL RESULTS

Simplified sketch of proof of 4.1. . Let $\mathbf{x}_{\mathrm{ID}}$ (resp. $\mathbf{x}_{\mathrm{OoD}}$) such that $\mathbf{x}_{\mathrm{ID}} \sim \mathbb{P}_{\mathrm{ID}}(\mathbf{x})$ (resp. $\mathbf{x}_{\mathrm{OoD}} \sim \mathbb{P}_{\mathrm{OoD}}(\mathbf{x})$). We suppose without loss of generality that for both inputs, softmax layer yields $y_\eta \in [\![1, K]\!]$ as a label and same cosine similarity.

Moreover we recall that $\mathbf{x}_{\mathrm{OoD}}$ is such that $\mathbf{x}_{\mathrm{OoD}}$ verify $(\mathbf{x}_{\mathrm{OoD}}, y_\eta) \in \{(\mathbf{x}, y) | \exists j \in 1, ...K, \cos(\phi_j) \geq 1 - \varepsilon\}$. Because MSP correctly classified $\mathbf{x}_{\mathrm{ID}}$ as ID and $\mathbf{x}_{\mathrm{OoD}}$ as OoD, it yields

$$\mathbb{P}(y_\eta | \mathbf{x}_{\mathrm{OoD}}) \propto \exp(\|\mathbf{w}_\eta\| \|\mathbf{z}_{\mathrm{OoD}}\| \cos(\phi_\eta)) \leq \mathbb{P}(y_\eta | \mathbf{x}_{\mathrm{ID}}) \propto \exp(\|\mathbf{w}_\eta\| \|\mathbf{z}_{\mathrm{ID}}\| \cos(\phi_\eta)). \tag{29}$$

Using that $x \mapsto \log(x)$ is increasing, it yields $\|\mathbf{z}_{\mathrm{OoD}}\| \leq \|\mathbf{z}_{\mathrm{ID}}\|$.
For a general proof, it should utilize the Neural Collapse (NC) property, which implies that all classification vectors have the same norm: $\|\mathbf{w}_1\| = \cdots = \|\mathbf{w}_K\|$. and should account for its angular property, which arises from the way $\mathbf{x}_{\mathrm{OoD}}$ is selected. □

Lemma C.1. *Let $V : \Omega \to \mathbb{R}^d$ be a random vector distributed uniformly on the spherical cap*

$$C_\Phi(\boldsymbol{\varphi}) \;=\; \{x \in S^{d-1} : x^\top \boldsymbol{\varphi} \geq \cos \Phi\},$$

where $\boldsymbol{\varphi} \in S^{d-1}$ is a fixed unit vector and $\Phi \in [0, \pi]$ is a given angle. Then, the first moment of V is given by

$$\mathbb{E}[V] \;=\; c\,\boldsymbol{\varphi},$$

Proof of 4.2. For all $i \in [\![1, M]\!]$ we set $\mathbf{z}^{(i)} = \|\mathbf{z}\|\boldsymbol{\alpha}^{(i)}$. Recall that $\mathrm{Var}(Z_M) = \mathrm{Tr}\left(\frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{z}^{(i)} - \boldsymbol{\mu})(\mathbf{z}^{(i)} - \boldsymbol{\mu})^T\right)$. Then :

$$\mathrm{Var}(Z_M) = \frac{1}{M-1}\sum_{i=1}^{M}\|\mathbf{z}^{(i)} - \boldsymbol{\mu}\|^2. \tag{30}$$

Expanding the sum knowing that $\forall i$, $\|\mathbf{z}_i\| = \|\mathbf{z}\|$ and $\boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M} z_i = \frac{\|\mathbf{z}\|}{M}\sum_{i=1}^{M}\boldsymbol{\alpha}^{(i)}$ we have:

$$\sum_{i=1}^{M}\|\mathbf{z}^{(i)} - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{M}\left(\|\mathbf{z}\|^2 - 2\|\mathbf{z}\|\langle\boldsymbol{\alpha}^{(i)}, \boldsymbol{\mu}\rangle + \|\boldsymbol{\mu}\|^2\right) = M\|\mathbf{z}\| - 2\frac{\|\mathbf{z}\|^2}{M}\langle\sum_{i=1}^{M}\boldsymbol{\alpha}^{(i)}, \sum_{j=1}^{M}\boldsymbol{\alpha}^{(j)}\rangle + M\|\boldsymbol{\mu}\|^2. \tag{31}$$

Expanding $\|\boldsymbol{\mu}\|^2 = \frac{\|\mathbf{z}\|^2}{M^2}\|\sum_{i=1}^{M}\boldsymbol{\alpha}^{(i)}\|^2 = \frac{\|\mathbf{z}\|^2}{M^2}\left(\sum_{i=1}^{M}\|\boldsymbol{\alpha}^{(i)}\|^2 + \sum_{i\neq j}\langle\boldsymbol{\alpha}^{(i)}, \boldsymbol{\alpha}^{(j)}\rangle\right)$, same for $\langle\sum_{i=1}^{M}\boldsymbol{\alpha}^{(i)}, \sum_{j=1}^{M}\boldsymbol{\alpha}^{(j)}\rangle$, we have :

$$\frac{1}{M-1}\sum_{i=1}^{M}\|\mathbf{z}^{(i)} - \boldsymbol{\mu}\|^2 = \frac{M}{M-1}\|\mathbf{z}\|^2 - \frac{\|\mathbf{z}\|^2}{M(M-1)}\left(\sum_{i=1}^{M}\|\boldsymbol{\alpha}^{(i)}\|^2 + \sum_{i\neq j}\langle\boldsymbol{\alpha}^{(i)}, \boldsymbol{\alpha}^{(j)}\rangle\right). \tag{32}$$

As $\boldsymbol{\alpha}^{(i)}, \ldots\boldsymbol{\alpha}^{(M)}$ are i.i.d., uniform on the spherical cap, we have $\mathbb{E}[\boldsymbol{\alpha}^{(i)}] = c\boldsymbol{\varphi}$, taking expectation, we have :

$$\mathbb{E}\left[\sum_{i=1}^{M}\|\boldsymbol{\alpha}^{(i)}\|^2 + \sum_{i\neq j}\langle\boldsymbol{\alpha}^{(i)}, \boldsymbol{\alpha}^{(j)}\rangle\right] = M + M(M-1)c^2. \tag{33}$$

Therefore we have the following **expectation equality**:

$$\mathbb{E}[\mathrm{Var}(Z_M)] = \frac{M\|\mathbf{z}\|^2}{M-1} - \frac{\|\mathbf{z}\|^2(1 + (M-1)c^2)}{M-1} = \|\mathbf{z}\|^2\left(\frac{M-1}{M-1} - \frac{(M-1)c^2}{M-1}\right) = (1 - c^2)\|\mathbf{z}\|^2. \tag{34}$$

Taking back Eq. (31) and observing that by the strong law of large numbers,

$$\|\boldsymbol{\mu}\|^2 = \frac{\|\mathbf{z}\|^2}{M^2}\|\sum_{i=1}^{M}\boldsymbol{\alpha}^{(i)}\|^2 \to_{a.s} c^2\|\mathbf{z}\|^2, \quad M \to +\infty. \tag{35}$$

Therefore, we have the **following asymptotic convergence**:

$$\mathrm{Var}(Z_M) \to_{a.s} (1 - c^2)\|\mathbf{z}\|^2, \quad M \to +\infty. \tag{36}$$

□

D CONCENTRATION INEQUALITY

In high dimensions, naive Monte Carlo and rejection sampling almost never cover the full geometry because concentration of measure confines nearly all the volume to a thin shell. This makes estimating uncertainty with a finite sample impractical and undermines the reliability of traditional measures. Instead, we embrace the curse of dimensionality by adopting a probabilistic framework that turns it into a blessing: rather than fight measure concentration, we exploit it to build our method.

To introduce our method—and for pedagogical clarity—we begin with the simplest case of a uniform distribution on the sphere, develop the necessary subgaussian theory, and discuss the locally uniform case. Only then do we specialize to the projected Gaussian setting that underlies our estimator. By leveraging measure concentration, we obtain a theoretically justified and efficient estimator of the true variance rather than attempting to reconstruct any arbitrary high-dimensional geometry.

Overall, our goal is to derive the general inequality 64, which establishes a rigorous and effective connection between high-dimensional probability and deep learning. It show how the phenomenon of concentration crucially depends on the structure of the embedding vector.

D.1 UNIFORM ASSUMPTION

Theorem D.1. *Let $\boldsymbol{\varphi} \in S^{D-1}$ and $\Phi \in [0, \pi]$ be given, and define the spherical cap:*

$$C_\Phi(\boldsymbol{\varphi}) \ = \ \Big\{\boldsymbol{\alpha} \in S^{D-1} : \boldsymbol{\alpha}^\top \boldsymbol{\varphi} \geq \cos\Phi\Big\}. \tag{37}$$

The concentration parameter $c = \mathbb{E}\big[\boldsymbol{\alpha}^\top \boldsymbol{\varphi}\big]$ quantifies how tightly the perturbed directions are distributed around $\boldsymbol{\varphi}$.

For $M \in \mathbb{N}^*$ we suppose $\{\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \dots \ \boldsymbol{\alpha}^{(M)}\}$ is a sequence of i.i.d. random vectors with the uniform distribution on $C_\Phi(\boldsymbol{\varphi})$. We define the matrix :

$$A_M := \Big(\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \dots, \boldsymbol{\alpha}^{(M)}\Big), \tag{38}$$

and we denote $\operatorname{Var}(A_M)$ its variance estimator. Then the following hold : As $M \to \infty$,

$$\operatorname{Var}(A_M) \ \xrightarrow{a.s.} \ \big(1 - c^2\big). \tag{39}$$

We are interested in understanding both how quickly $\operatorname{Var}(A_M)$ converges to $(1 - c^2)$ when $M \to +\infty$.

For the sake of simplicity, **we suppose** $\forall i, \boldsymbol{\alpha}^{(i)}$ follow an uniform distribution over all the unit hypersphere S^{D-1} such that $c^2 = 0$. We also define the following perturbed estimator in its case :

$$\widetilde{\operatorname{Var} A_M^f} := \operatorname{Tr}\left(\frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{\alpha}^{(i)} - \boldsymbol{\mu})(\boldsymbol{\alpha}^{(i)} - \boldsymbol{\mu})^T\right), \tag{40}$$

where $\boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\alpha}^{(i)}$

Since $\frac{M}{M-1} \simeq 1$ for $M \geq 10$, this approximation does not affect practical computations and is used solely for convenience.

Theorem D.2. *(Uniform on the whole sphere) For all $\epsilon > 0$, we have*

$$\mathbb{P}\Big(1 - \widetilde{\operatorname{Var} A_M^f} > \epsilon\Big) \leq \exp\Big(-c_1\, D\, \epsilon\, M\Big). \tag{41}$$

where $c_1 > 0$ is an absolute constant.

Theorem D.3. *(Locally uniform on spherical cap) For all $\epsilon > 0$, we have*

$$\mathbb{P}\Big((1 - c^2) - \widetilde{\operatorname{Var} A_M} > \epsilon\Big) \leq \exp\Big(-c_1\, D\, \epsilon\, M\Big). \tag{42}$$

where $c_1 > 0$ is an absolute constant.

To prove the Theorem.D.2, we first need to explore some key properties of sub-gaussian vectors.

D.1.1 Preliminaries on sub-gaussian vectors

Definition D.1 (Sub-gaussian random variable)**.** *We say that real random variable X is sub-gaussian if there is a constant $C > 0$ such that for $t \geq 0$:*

$$\mathbb{P}(|X| > t) \leq 2\exp(-t^2/C^2), \tag{43}$$

Its subgaussian norm is the quantity

$$\|X\|_{\psi_2} = \inf_{\lambda>0} \mathbb{E}\left[\exp\left(\frac{X^2}{\lambda^2}\right) \leq 2\right]. \tag{44}$$

Definition D.2 (Sub-gaussian random vector)**.** *We say that the random vector X is sub-gaussian if and only if*

$$\|X\|_{\psi_2} := \sup_{u\in S^{D-1}} \|u^T X\|_{\psi_2} < +\infty. \tag{45}$$

Lemma D.4. *If $\boldsymbol{\alpha}$ is a random vector following a uniform distribution on the unit hypersphere S^{D-1} then $\boldsymbol{\alpha}$ is sub-gaussian such that $\|\boldsymbol{\alpha}\|_{\psi_2} = O(\frac{1}{\sqrt{d}})$.*

Lemma D.5. *Let $\boldsymbol{\alpha}^{(1)}, \dots \boldsymbol{\alpha}^{(M)}$ be M random vector i.i.d. following an uniform distribution on the unit hypersphere. Let $S := \sum_{i=1}^{M} \boldsymbol{\alpha}^{(i)}$. Then $\mathbb{E}[S] = 0$ and is sub-gaussian with $\|S\|_{\psi_2} = O(\frac{\sqrt{M}}{\sqrt{D}})$.*

Proof. Let $u \in S^{D-1}$. Then:

$$\|S\|_{\psi_2} = \|u^T \sum_{i=1}^{M} \boldsymbol{\alpha}^{(i)}\|_{\psi_2} = \|\sum_{i=1}^{M} u^T \boldsymbol{\alpha}^{(i)}\|_{\psi_2} \leq K \left(\sum_{i=1}^{M} \|u^T \boldsymbol{\alpha}^{(i)}\|_{\psi_2}\right)^{1/2} \leq K\frac{\sqrt{M}}{\sqrt{D}}, \tag{46}$$

where $K > 0$ is an absolute constant. A complete proof of the penultimate inequality is provided in Vershynin [2018]. □

Corollary D.5.1. $\forall \lambda > 0$ *and* $u \in \mathbb{R}^D$*,*

$$\mathbb{E}[\exp \lambda u^T S] \leq \exp\left(\frac{K^2 M}{2D} \lambda^2 \|u\|^2\right), \tag{47}$$

where $K > 0$ is an absolute constant proportional to the subgaussian norm of S.

The following theorem gives us a concentration inequality on $\|S\|^2$. We consider it as a "weak" version of the Hanson-Wright inequality since it does not require the components of the vector to be independent but subgaussian with the cost of A to be positive- semidefinite. We write the inequality as it is in Hsu et al. [2012].

Theorem D.6 (Weak Hanson-Wright Inequality)**.** *Suppose that a random vector $X \in \mathbb{R}^n$ satisfies*

$$\mathbb{E}\exp(u^T(X - \boldsymbol{\eta})) \leq \exp\Bigl(\frac{\sigma^2\|u\|^2}{2}\Bigr), \forall u \in \mathbb{R}^n. \tag{48}$$

Then, for any definite positive matrix $A \in \mathbb{R}^{m\times n}$ (with $\Sigma = A^T A$), for all $t > 0$:

$$\mathbb{P}\Bigl(\|AX\|^2 > \sigma^2\Bigl(\operatorname{tr}(\Sigma) + 2\sqrt{\operatorname{tr}(\Sigma^2)\, t} + 2\|\Sigma\|\, t\Bigr)\Bigr) \leq e^{-t}. \tag{49}$$

where $\|\Sigma\|$ denotes for the spectral (operator) norm of the matrix Σ.

Proof. A complete proof of this theorem is provided in Hsu et al. [2012]. □

D.1.2 Proof of Theorem. D.2

Proof. Recall (32), then replace $M-1$ by M to the denominator, we obtain

$$\widetilde{\operatorname{Var} A_M} = 1 - \|\boldsymbol{\mu}\|^2 \quad \text{(where } \|\mathbf{z}\| = 1). \tag{50}$$

Our goal is to bound the deviation $\widetilde{\operatorname{Var} A_M} - 1$ which reduces to control $\|\boldsymbol{\mu}\|^2$. Let $S = \sum_{i=1}^{M} \boldsymbol{\alpha}^{(i)}$. Then :

$$\|\boldsymbol{\mu}\|^2 = \frac{\|S\|^2}{M^2}. \tag{51}$$

Thus, the error in the variance estimation is

$$|1 - \widetilde{\operatorname{Var} A_M}| = 1 - \widetilde{\operatorname{Var} A_M} = \frac{\|S\|^2}{M^2}. \tag{52}$$

We apply the Weak Hanson-Wright Inequality D.6 and the Corollary. D.5.1 We take $\boldsymbol{\eta} = 0$, and $A = I_D$, so that $\Sigma = I_D$, with

$$\operatorname{Tr}(I_D) = D, \quad \operatorname{Tr}(I_D^2) = D, \quad \|I_D\| = 1 \quad \sigma^2 = K^2\,\frac{M}{D}.$$

It follows that we have that with probability at least $1 - e^{-t}$,

$$\|S\|^2 \leq C^2\,M\Big(1 + 2\sqrt{\frac{t}{D}} + 2\frac{t}{D}\Big). \tag{53}$$

and it follows that :

$$\|\boldsymbol{\mu}\|^2 \leq \frac{K^2}{M^2}\;M\Big(1 + 2\sqrt{\frac{t}{D}} + 2\frac{t}{D}\Big) = \frac{K^2}{M}\Big(1 + 2\sqrt{\frac{t}{D}} + 2\frac{t}{D}\Big). \tag{54}$$

Thus, for every $t \geq 0$ to get,

$$1 - \widetilde{\operatorname{Var}(A_M)} = \|\boldsymbol{\mu}\|^2 \leq \epsilon, \tag{55}$$

it is sufficient that:

$$\Big(1 + 2\sqrt{\frac{t}{D}} + 2\frac{t}{D}\Big) \leq \frac{M\epsilon}{K^2}. \tag{56}$$

Suppose that M is large enough such that the right term is greater than 1, then we have $c_1 > 0$ such that:

$$t \geq c_1\,D\,\epsilon\,M, \tag{57}$$

can be chosen such as $\|\boldsymbol{\mu}\|^2 \leq \epsilon$ holds.

Thus, we deduce that:

$$\mathbb{P}\Big(1 - \widetilde{\operatorname{Var}(A_M)} > \epsilon\Big) = \mathbb{P}\Big(\|\boldsymbol{\mu}\|^2 > \epsilon\Big) \leq \exp\Big(-c_1\,D\,\epsilon\,M\Big).$$

□

The general case stated in Theorem D.3 can be proved by combining advanced results on manifolds with convex boundaries, the log-Sobolev inequality, and the Bakry—Émery criterion. First, one shows that for any $\|f\|_{Lip}$-Lipschitz function F and any random vector X uniformly distributed on a spherical cap, the following concentration estimate holds:

$$\mathbb{P}\left(|F(\mathbf{X}) - E[F(\mathbf{X})]| > r\right) \leq \exp\left(-\frac{CDr^2}{\|f\|_{Lip}^2}\right) \tag{58}$$

where C is an absolute constant. The proof is beyond the scope of this work and can be found in Kolesnikov and Milman [2016], Ledoux [2001].

D.2 PROJECTED GAUSSIAN ASSUMPTION

The projected Gaussian assumption covers the general case of measure concentration for Lipschitz functions on $\mathbb{R}^D$, relates precisely to the normalization process in Alg. 2 and Alg. 3. Its behavior is well understood using log-Sobolev and Herbst inequalities: its rigorous justification is also relying on log–Sobolev Inequality and Herbst arguments, and full proof of the following results can be found in [Ledoux, 2001].

Theorem D.7. *Suppose that for all $F : \mathbb{R}^D \to \mathbb{R}$ Lipschitz, the law of the random variable $F(\mathbf{X})$ verify Log-Sobolev Inequality assumption, in the sense that for all $r \geq 0$,*

$$\mathbb{P}(|F(\mathbf{X}) - \mathbb{E}[F(\mathbf{X})]| \geq r) \leq c\exp\left(-\frac{r^2}{2C\|F\|_{Lip}^2}\right), \tag{59}$$

for some absolute constants $c, C > 0$. Then for all $F : \mathbb{R}^D \to \mathbb{R}$ Lipschitz and $r \geq \sigma \|F\|_{Lip}$,

$$\mathbb{P}\left(\left|F\left(\frac{\mathbf{X}}{|\mathbf{X}|}\right) - \mathbb{E}\left[F\left(\frac{\mathbf{X}}{|\mathbf{X}|}\right)\right]\right| \geq r\right) \leq 2c \exp\left(-\frac{\eta^2}{8C}\left(\frac{r}{\|F\|_{Lip}} - \sigma\right)^2\right), \tag{60}$$

where

$$\eta := \mathbb{E}[|\mathbf{X}|] \quad \text{and} \quad \sigma := \mathbb{E}\left[\left|\frac{|\mathbf{X}|}{\eta} - 1\right|\right]. \tag{61}$$

The quantity $\|F\|_{\text{Lip}} := \sup_{\mathbf{x},\mathbf{y} \in \mathbb{R}^D : \mathbf{x} \neq \mathbf{y}} \frac{|F(\mathbf{x}) - F(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|}$ is the Lipschitz norm of F with respect to the Euclidean norm on $\mathbb{R}^D$.

Corollary D.7.1. *Consider the Gaussian case $X \sim \mathcal{N}(m, \Sigma)$, $m \in \mathbb{R}^D$, $\Sigma \in \text{Sym}^+_{D \times D}(\mathbb{R})$. Then the sub-Gaussian concentration of Lipschitz functions holds with*

$$c = 2 \quad \text{and} \quad C = \|\Sigma\|_{\text{op.}} = \max_{|x|=1} \langle \Sigma x, x \rangle. \tag{62}$$

Moreover, one can show that $\mathbb{E}\|X\| = K'D$, leading to a concentration inequality that depends on the dimensionality D:

$$\mathbb{P}\left(\left|F\left(\frac{\mathbf{X}}{\|\mathbf{X}\|}\right) - \mathbb{E}\left[F\left(\frac{\mathbf{X}}{\|\mathbf{X}\|}\right)\right]\right| \geq r\right) \leq 2c \exp\left(-\frac{KDr^2}{\|F\|_{\text{Lip}}}\right), \tag{63}$$

where K is an absolute constant depending on K', σ and C.

D.3 MEASURE TENSORIZATION

Considering $\mathbf{Y}_1, ...\mathbf{Y}_M$ i.i.d such as $\forall i \in [|1, M|], \mathbf{Y_i}$ follow the same distribution as $\frac{\mathbf{X}}{\|\mathbf{X}\|}$. Because the Log-Sobolev Inequality is stable by measure tensorization [Chafaï and Lehec, 2024] and because the variance operator is an Lipschitz function on the unit hypersphere, it follow that if $\text{Var}\left(\mathbf{Y}_1, ...\mathbf{Y}_M\right)$ denotes the empirical variance:

$$\mathbb{P}\left(\left|\text{Var}\left(\mathbf{Y}_1, ...\mathbf{Y}_M\right)\right) - \text{Var}\left(\frac{\mathbf{X}}{\|\mathbf{X}\|}\right)\right| \geq r\right) \leq 2c \exp\left(-MKr^2D\right) \tag{64}$$

where K is an absolute constant depending on K', σ, C and the Lipchitz constant.

In our method, the normalized output of the DC Layer $\forall i \in [|1, M|]$, $\frac{\boldsymbol{\alpha}_{DC}^{(i)}}{\|\boldsymbol{\alpha}_{DC}^{(i)}\|} := \tilde{\boldsymbol{\alpha}}^{(i)}(\boldsymbol{z})$ follows a projected Gaussian distribution and its true variance denoted here by $\text{Var}(\boldsymbol{\alpha_{DC}}(\boldsymbol{z}))$ depends on the embedding z. The absolute constant K depends on the embedding z, hence we write $K = K(z)$. It follows that:

$$\mathbb{P}\Big(\big|\text{Var}(\tilde{\boldsymbol{\alpha}}^{(1)}(\boldsymbol{z}), \ldots, \tilde{\boldsymbol{\alpha}}^{(M)}(\boldsymbol{z})) - \text{Var}(\boldsymbol{\alpha_{DC}}(\boldsymbol{z}))|\big) \leq 2c \exp\big(-MK(z)r^2D\big). \tag{65}$$

This inequality also holds in the uniform case, as this setup also satisfies the Log-Sobolev inequality, which permits a measure tensorization argument.

In practice, for OoD data, one expects the empirical variance to converges more slowly—both because it has intrinsically higher variance and because OoD examples can be highly diverse—so the required number of samples M must be only calibrated on a ID validation set.

Relying on the theoretical observations established by Sun et al. [2021], one can easily show that the constant $K(z)$ is larger for ID embeddings than for OoD embeddings, reflecting the fact that ID embeddings concentrate more tightly in the representation space and exhibit lower variance than OoD embeddings. A more precise study of this constant is left for future work.