
Double Descent Meets Out-of-Distribution Detection: Theoretical Insights and Empirical Analysis on the Role of Model Complexity

Mouïin Ben Ammar^{◊, †, *}, David Brellmann^{†, *}, Arturo Mendoza[†], Antoine Manzanera[◊], Gianni Franchi[◊]
U2IS Lab ENSTA Paris[◊], Palaiseau, FRANCE
Safran Tech[†], Chateaufort 78117, FRANCE
{first.last}@ensta.fr[◊], safrangroup.com[†]

Abstract

Out-of-distribution (OOD) detection is essential for ensuring the reliability and safety of machine learning systems. In recent years, it has received increasing attention, particularly through post-hoc detection and training-based methods. In this paper, we focus on **post-hoc OOD detection**, which enables identifying OOD samples without altering the model’s training procedure or objective. Our primary goal is to investigate the relationship between **model capacity** and its OOD detection performance. Specifically, we aim to answer the following question: *Does the Double Descent phenomenon manifest in post-hoc OOD detection?* This question is crucial, as it can reveal whether overparameterization, which is already known to benefit generalization, can also enhance OOD detection. Despite the growing interest in these topics by the classic supervised machine learning community, this intersection remains unexplored for OOD detection. We empirically demonstrate that the Double Descent effect does indeed appear in post-hoc OOD detection. Furthermore, we provide theoretical insights to explain why this phenomenon emerges in such setting. Finally, we show that the overparameterized regime does not yield superior results consistently, and we propose a method to identify the optimal regime for OOD detection based on our observations. Code available [here](#)

1 Introduction

Since the breakthrough of AlexNet in 2012 (Krizhevsky et al., 2012), deep learning has seen rapid progress and has become the foundation for solving a wide variety of complex tasks across domains such as vision, language, and robotics (LeCun et al., 2015; Bengio, 2009). While Occam’s Razor suggests favoring simpler models with fewer parameters (Blumer et al., 1987), deep neural networks (DNNs) with massive overparameterization have nonetheless demonstrated remarkable generalization ability in practice.

Traditionally, generalization performance has been characterized by a U-shaped test error curve with respect to model complexity (Geman et al., 1992; Hastie, 2009), recommending a “sweet spot” where the model is expressive enough to avoid underfitting but not so complex that it overfits. However, this classical view has been challenged in 2019 by the emergence of the *double descent* phenomenon (Belkin et al., 2019), which reveals a second descent in test error after the interpolation threshold. This insight has reshaped our understanding of the generalization behavior of overparameterized models.

Generalization typically refers to in-distribution (ID) performance—where test data are assumed to be drawn i.i.d. from the same distribution as training data. However, this closed-world assump-

*Equal contribution

tion (Scheirer et al., 2012) rarely holds in real-world scenarios. In open-world settings (Bendale & Boulton, 2016; Drummond & Shearer, 2006), models frequently encounter *out-of-distribution* (OOD) inputs that differ significantly from training data, either due to *semantic shift* (e.g., new classes) (Hendrycks & Gimpel, 2017) or *covariate shift* (e.g., domain changes) (Ben-David et al., 2010; Wang & Deng, 2018; Li et al., 2017).

In this work, we focus on OOD detection under semantic shifts as defined in Yang et al. (2024), where the goal is to identify inputs from unseen classes. More precisely, we investigate *post-hoc* OOD detection methods (DeVries & Taylor, 2018; Liu et al., 2020; Sun et al., 2021), which operate on top of a pretrained classifier without altering its training process. These methods are widely used due to their practicality and modularity.

Our central research question is: **How does model capacity affect the effectiveness of post-hoc OOD detection?** While double descent has been studied in the context of ID generalization, its role in OOD detection remains unexplored. We empirically demonstrate that post-hoc OOD detection also exhibits a double descent curve with respect to model complexity. This observation is important as it suggests that overparameterization is not universally beneficial for OOD detection.

We conduct extensive experiments across different architectures (CNNs, ResNets, ViTs, Swin) and post-hoc detection methods to confirm this trend. Furthermore, under a simplified Gaussian mixture model, we derive theoretical insights using tools from random matrix theory that mirror the empirical findings. Surprisingly, we also observe cases where the underparameterized regime achieves better OOD detection performance, prompting us to investigate these cases through the lens of Neural Collapse (Papayan et al., 2020), which provides a geometric interpretation of the learned representations near convergence.

Contributions. We advance the theoretical and empirical understanding of model complexity in post-hoc OOD detection:

1. We are the first to empirically highlight a double descent phenomenon in OOD detection across both CNN and transformer architectures.
2. We provide theoretical insight using random matrix theory, showing that both ID and OOD risks peak at the interpolation threshold.
3. We show that overparameterization is not always optimal for OOD detection, and propose a Neural Collapse-based criterion to detect when simpler models may perform better.

2 Related work

OOD Detection. OOD detection research focuses on two primary directions: Training-based and post-hoc methods. We will focus on latter ones. These can be categorized based on the essential feature used for the scoring function. First, logit- or confidence-based methods leverage network logits to derive a confidence measure used as an OOD scoring metric (Hendrycks & Gimpel, 2017; DeVries & Taylor, 2018; Liu et al., 2020; Huang & Li, 2021; Hendrycks et al., 2022). A common

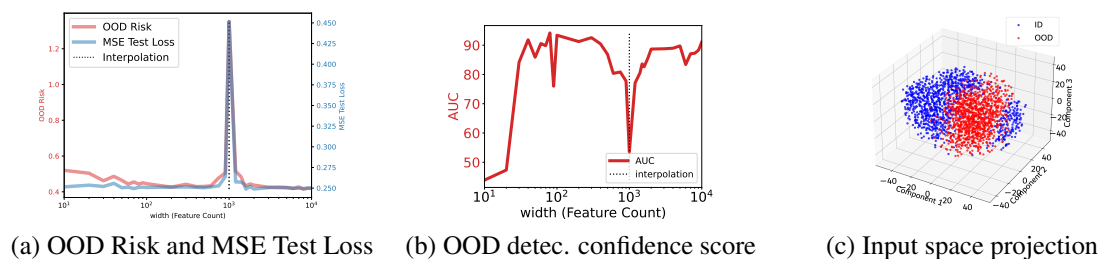


Figure 1: Illustration of the double descent phenomenon in a Random ReLU Feature model as a function of model width (log scale). **(a)** Evolution of the in-distribution (ID) Mean-Squared Error (MSE) and OOD detection risk. **(b)** Confidence-based OOD score defined in equation 14. **(c)** Three-dimensional t-SNE projection of the input space, visualizing the separation between ID and OOD samples. The ID samples ($n = 1,000$) are drawn from a Gaussian Mixture Model (GMM) fitted to a subset of CIFAR-10, while OOD samples are drawn from an independent GMM.

baseline for this methods is the softmax score (Hendrycks & Gimpel, 2017), which simply uses the model softmax prediction as the OOD score. Then, Energy (Liu et al., 2020) elaborates on it by computing the LogSumExp on the logits, thus offering empirical and theoretical advantages over the Softmax confidence score. Second, feature-based and hybrid methods (Lee et al., 2018b; Wang et al., 2022; Sun et al., 2022; Ming et al., 2023; Djuricic et al., 2023; Ammar et al., 2024) exploit the model’s final representation to derive the scoring function. Hybrid methods further augments the scoring metrics defined on the features using logits as weighting factors. Mahalanobis (Lee et al., 2018b) estimates density on ID training samples using a mixture of class-conditional Gaussians. NECO (Ammar et al., 2024), on the contrary, leverages the geometric properties of Neural Collapse to construct a scoring function based on the relative norm of a sample within the subspace defined by the ID data. Besides OOD detection methods, Fang et al. (2024) have used risk formulations closely related to the one we use in this work to study PAC bounds for OOD detection.

Double Descent. The double descent risk curve was introduced by Belkin et al. (2019) to explain the good performance observed in practice by overparameterized models (Zhang et al., 2021; Belkin et al., 2018; Nakkiran et al., 2021) and to bridge the gap between the classical bias-variance trade-off theory and modern practices. Theoretical investigation into this phenomenon mainly focuses on various linear models in both regression and classification problems through the Random Matrix Theory (Louart et al., 2018; Liao et al., 2020; Jacot et al., 2020; Dereziński et al., 2020; Kini & Thrampoulidis, 2020; Mei & Montanari, 2022; Deng et al., 2022; Bach, 2024; Brellmann et al., 2024), techniques from statistical mechanics (D’Ascoli et al., 2020; Canatar et al., 2021), the VC theory (Lee & Cherkassky, 2022; Cherkassky & Lee, 2024), or novel bias-variance decomposition in deep neural networks (Yang et al., 2020). The double descent phenomenon has also been observed in experiments with popular neural network architectures (Belkin et al., 2019; Nakkiran et al., 2021). In addition to depending on the model complexity, the double descent phenomenon also depends on other dimensions such as the level of regularization (Liao et al., 2020; Mei & Montanari, 2022), the number of epochs (Nakkiran et al., 2021; Stephenson & Lee, 2021; Olmin & Lindsten, 2024), or the data eigen-profile (Liu et al., 2021a). Finally, the theoretical background on double descent and benign overparametrization developed by Bartlett et al. (2020) inspired subsequent works that focused on generalization under dataset shifts (Tripuraneni et al., 2021b; Hao et al., 2024; Kausik et al., 2024; Hao & Zhang, 2024). It is important to note that these dataset shifts concern scenarios where the model can generalize, *e.g.*, the class labels are the same as in the training. These studies do not address OOD detection.

3 Preliminaries

Notations. For a real vector $\mathbf{v}$, we denote by $\|\mathbf{v}\|_2$ the Euclidean norm of $\mathbf{v}$. When the matrix $\mathbf{A}$ is full rank, we denote by $\mathbf{A}^+$ the Moore-Penrose inverse of $\mathbf{A}$. We depict by $[d] := \{1, \dots, d\}$ the set of the d first natural integers. For a subset $\mathcal{T} \subseteq [d]$, we denote by $\mathcal{T}^c := [d] \setminus \mathcal{T}$ its complement set. For a subset $\mathcal{T} \subseteq [d]$, a d -dimensional vector $\mathbf{v} \in \mathbb{R}^d$ and a $n \times d$ matrix $\mathbf{A} = [\mathbf{a}^{(1)} | \dots | \mathbf{a}^{(n)}]^T \in \mathbb{R}^{n \times d}$, we use $\mathbf{v}_{\mathcal{T}} = [\mathbf{v}_j : j \in \mathcal{T}]$ to denote its $|\mathcal{T}|$ -dimensional subvector of entries from $\mathcal{T}$ and $\mathbf{A}_{\mathcal{T}} = [\mathbf{a}_{\mathcal{T}}^{(1)} | \dots | \mathbf{a}_{\mathcal{T}}^{(n)}]^T$ to denote the $n \times |\mathcal{T}|$ design matrix with variables from $\mathcal{T}$. For a subset $\mathcal{T} \subseteq [d]$, we define $R^d(\mathcal{T}) := \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}_{\mathcal{T}^c} = \mathbf{0}_{d-|\mathcal{T}|\}$. We use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to depict the min and max eigenvalues of $\mathbf{A}$, respectively. $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ denotes the standard multivariate Gaussian distribution of d random variables.

Supervised Learning Problems. In supervised learning, we use training dataset $\mathcal{D} := \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ of n independent and identically distributed (i.i.d.) samples drawn from an unknown distribution $P_{\mathcal{X}, \mathcal{Y}}$ over $\mathcal{X} \times \mathcal{Y}$. Using samples from the dataset $\mathcal{D}$, the objective is to find a predictor $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$ among a class of functions $\mathcal{F}$ to predict the target $y \in \mathcal{Y}$ of a new sample $\mathbf{x} \in \mathcal{X}$. In particular, given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, the objective is to minimize the expected risk (or loss) defined, for all $\hat{f} \in \mathcal{F}$, as:

$$R(\hat{f}) = \mathbb{E}_{(\mathbf{x}, y) \sim P_{\mathcal{X}, \mathcal{Y}}} [\ell(\hat{f}(\mathbf{x}), y)]. \quad (1)$$

Typically, we choose the mean-squared loss $\ell(\hat{f}(\mathbf{x}), y) = (\hat{f}(\mathbf{x}) - y)^2$ for regression problems or the zero-one loss $\ell(\hat{f}(\mathbf{x}), y) = \mathbb{1}_{\hat{f}(\mathbf{x}) \neq y}$ for classification problems. We denote the optimal predictor

by $f^* := \arg \min_{f \in \mathcal{F}} R(f)$. Since the distribution $P_{\mathcal{X}, \mathcal{Y}}$ is unknown in practice, we instead try to minimize an empirical version of the expected risk based on the dataset $\mathcal{D} := \{(\mathbf{x}_i, y_i)\}_{i=1}^n$:

$$R_{emp}(\hat{f}) = \frac{1}{n} \sum_{i=1}^n \ell(\hat{f}(\mathbf{x}_i), y_i). \quad (2)$$

Out-of-Distribution Detection. In machine learning problems, we usually assume that the test data distribution is similar to the training data distribution (the closed-world assumption). As this is not the case in real-world applications, the Out-of-Distribution (OOD) detection (Yang et al., 2024) aims to flag inputs that significantly deviate from the training data to prevent unreliable predictions. In the following, we denote by $P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}$ a distribution over $\mathcal{X} \times \mathcal{Y}$ that differs from the training distribution $P_{\mathcal{X}, \mathcal{Y}}$. As proposed in (Yang et al., 2024) we consider OOD that involves a semantic shift and represent concepts or labels not seen during training. A popular class of OOD detection techniques relies on the definition of a scoring function $s(\cdot; \hat{f}) = \hat{f}(\cdot)$, which uses the probability predictions of the classifier $\hat{f}(\cdot)$ as scores to flag an instance $\mathbf{x}$ as OOD when the score probability prediction is below a certain threshold λ . For example, a common approach is to use the Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017) that returns the higher softmax probabilities of the predictor $\hat{f}(\cdot)$ as a scoring function to measure the prediction confidence.

4 Insights on the Double Descent for the Binary Classification in Gaussian Covariate Model

In this section, we introduce the expected OOD risk metric and we present our main theoretical results on binary least-squares classifiers applied to Gaussian data. We assume that $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} := [0, 1]$. Let $\phi: \mathbb{R} \rightarrow \mathcal{Y}$ be a mapping, we denote by $\mathcal{F}_d := \{f: \mathcal{X} \rightarrow \mathcal{Y}, \mathbf{x} \mapsto \phi(\mathbf{x}^T \mathbf{w}) \mid \mathbf{w} \in \mathbb{R}^d\}$ the class of functions considered in this study.

4.1 System Model

Gaussian Covariate Model & Binary Classification. Let $\mathbf{w}^* \in \mathbb{R}^d$ and let $f^*: \mathbf{x} \mapsto \phi(\mathbf{x}^T \mathbf{w}^*)$ be an optimal binary classifier. We assume we have a training dataset $\mathcal{D} := \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of n i.i.d samples drawn from a Gaussian covariate model, *i.e.*, from a distribution $P_{\mathcal{X}, \mathcal{Y}}$ over $\mathcal{X} \times \mathcal{Y}$; where $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $y_i = \phi(z_i) = \phi(\mathbf{x}_i^T \mathbf{w}^* + \epsilon_i)$ is the noisy response of $\mathbf{x}_i$ with respect to $f^*(\cdot)$ in which $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a noise capturing approximation errors on logits z_i with $\sigma > 0$. In binary classification problems, the objective is to find a classifier $\hat{f}(\cdot) \in \mathcal{F}_d$ that fits $f^*(\cdot)$ using the training dataset $\mathcal{D}$. Although simple, the Gaussian covariate model is also considered in the double descent literature to provide theoretical insights (Belkin et al., 2020; Mei & Montanari, 2022).

Least-Squares Binary Classifiers. In this section, we consider least-squares binary classifiers to approximate logits of the optimal binary classifier $f^*(\cdot)$. Specifically, we define the least-squares binary classifier $\hat{f}: \mathbf{x} \mapsto \phi(\mathbf{x}^T \hat{\mathbf{w}})$ in which $\hat{\mathbf{w}}$ is obtained by analytically solving:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} R_{emp}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^T \mathbf{w} - z_i)^2 = \frac{1}{n} \|\mathbf{X} \mathbf{w} - \mathbf{z}\|_2^2, \quad (3)$$

where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ is the data matrix containing the n samples $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{z} = [z_1, \dots, z_n]^T \in \mathbb{R}^n$ is the target vector of noisy logits. We consider a particular subset of least-squares binary classifier $\hat{f}_{\mathcal{T}} \in \mathcal{F}_d$ that uses a subset $\mathcal{T} \subseteq [d]$ of p features that fits coefficients $\hat{\mathbf{w}} \in \mathbb{R}^d(\mathcal{T})$ as

$$\hat{\mathbf{w}}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^+ \mathbf{z} \quad \text{and} \quad \hat{\mathbf{w}}_{\mathcal{T}^c} = \mathbf{0}_{d-p}. \quad (4)$$

Out-of-Distribution Risk. To measure the ability of binary classifiers $\hat{f}(\cdot) \in \mathcal{F}_d$ to provide prediction confidence on samples drawn from both the training distribution $P_{\mathcal{X}, \mathcal{Y}}$ and the OOD distribution $P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}$, we introduce an expected OOD risk function similar to the expected risk defined in equation 3. Let $f^{\text{OOD}}: \mathbf{x} \mapsto \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}}) \in \mathcal{F}_d$ be an optimal classifier such that $f^{\text{OOD}}(\mathbf{x})$ is

close to 0.5 when the sample $\mathbf{x}$ is more likely drawn from the $P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}$ and close to $f^*(\mathbf{x})$ when $\mathbf{x}$ is more likely drawn from $P_{\mathcal{X},\mathcal{Y}}$. We define the expected OOD risk $R_{\text{OOD}} : \mathcal{F}_d \rightarrow \mathbb{R}$ as:

$$R_{\text{OOD}}(\hat{f}) = \mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}} [(\hat{f}(\mathbf{x}) - f^{\text{OOD}}(\mathbf{x}))^2] + \mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}} [(\hat{f}(\mathbf{x}) - f^{\text{OOD}}(\mathbf{x}))^2], \quad (5)$$

which depicts the expected risk of the binary classifier $\hat{f}(\cdot)$ on the loss function $\ell : (\hat{y}, y) \mapsto (\hat{y} - y)^2$ and distributions $P_{\mathcal{X},\mathcal{Y}}$ and $P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}$.

Remark 4.1. A low value for $R_{\text{OOD}}(\hat{f})$ indicates two aspects: (i) the classifier $\hat{f}(\cdot)$ is confident on predictions over the distribution $P_{\mathcal{X},\mathcal{Y}}$, which corresponds to a low $\mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}} [(\hat{f}(\mathbf{x}) - f^{\text{OOD}}(\mathbf{x}))^2]$; and/or (ii) $\hat{f}(\cdot)$ is not confident on predictions over the distribution $P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}$, which is reflected by a low $\mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}} [(\hat{f}(\mathbf{x}) - f^{\text{OOD}}(\mathbf{x}))^2]$. In particular, $R_{\text{OOD}}(\hat{f})$ is minimized when logits are maximally confident on ID samples (only one logit is non-zero) and uniformly distributed on OOD samples.

Remark 4.2. Note that the expected OOD risk defined in equation 5 can be extended to multi-class classifiers $\hat{f}(\cdot)$ using the softmax function with

$$R_{\text{OOD}}(\hat{f}) = \mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}} [(\|\hat{f}(\mathbf{x})\|_{\infty} - \|f^{\text{OOD}}(\mathbf{x})\|_{\infty})^2] + \mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}} [(\|\hat{f}(\mathbf{x})\|_{\infty} - \|f^{\text{OOD}}(\mathbf{x})\|_{\infty})^2],$$

where $f^{\text{OOD}}(\mathbf{x})$ is close to $1/C$ when the sample $\mathbf{x}$ is more likely drawn from the $P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}$, C depicts the number of classes, and $\|\cdot\|_{\infty}$ denotes the infinity norm.

In order to use the Random Matrix Theory, we make the following assumptions.

Assumption 4.1. The activation function $\phi(\cdot)$ is strictly monotonically increasing. Furthermore, its derivative $\phi'(\cdot)$ is strictly positive and bounded.

Remark 4.3. This assumption holds for many of the activation functions traditionally considered in neural networks, such as sigmoid functions.

Assumption 4.2. Let $\Sigma, \Sigma^{\text{OOD}} \in \mathbb{R}^{d \times d}$ defined as

$$\Sigma = \mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}} \left[\left(\frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}})}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^{\text{OOD}}} \right)^2 \mathbf{x} \mathbf{x}^T \right] \text{ and } \Sigma^{\text{OOD}} = \mathbb{E}_{(\mathbf{x},\cdot) \sim P_{\mathcal{X},\mathcal{Y}}^{\text{OOD}}} \left[\left(\frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}})}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^{\text{OOD}}} \right)^2 \mathbf{x} \mathbf{x}^T \right].$$

We assume that

$$\arg \min_{\mathbf{a} \in \mathbb{R}^d} \frac{\mathbf{a}^T \Sigma \mathbf{a}}{\mathbf{a}^T \mathbf{a}} \notin R^d(\mathcal{T}) \quad \text{and} \quad \arg \min_{\mathbf{a} \in \mathbb{R}^d} \frac{\mathbf{a}^T \Sigma^{\text{OOD}} \mathbf{a}}{\mathbf{a}^T \mathbf{a}} \notin R^d(\mathcal{T}).$$

4.2 Out-Of-Distribution Risk

Leveraging the Random Matrix Theory and extending the Theorem 1 in Belkin et al. (2020), we can derive bounds for the expected risk (Appendix A.1) and for the expected OOD risk (proof in Appendix A.2) of the subset of classifiers defined with equation 4:

Theorem 1. Let $(p, q) \in [d]^2$ such that $p + q = d$, $\mathcal{T} \subseteq [d]$ with $|\mathcal{T}| = p$ an arbitrary subset of the d first natural integers, and $\mathcal{T}^c := [d] \setminus \mathcal{T}$ its complement set. Let $\hat{\mathbf{w}} \in \mathbb{R}^d(\mathcal{T})$ such that $\hat{\mathbf{w}}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^+ \mathbf{z} \in \mathbb{R}^p$ and $\hat{\mathbf{w}}_{\mathcal{T}^c} = \mathbf{0}_q \in \mathbb{R}^q$. Under Assumptions 4.1-4.2, the expected OOD risk on the predictor $\hat{f}_{\mathcal{T}} : \mathbf{x} \mapsto \phi(\mathbf{x}^T \hat{\mathbf{w}})$ satisfies

$$c \, c(n, p) \leq \mathbb{E}_{\mathbf{X}} [R_{\text{OOD}}(\hat{f})] \leq C \, c(n, p),$$

where $c, C > 0$ and

$$c(n, p) = \begin{cases} \frac{p}{n-p-1} \left(\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2 \right) + \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 & \text{if } p \leq n-2, \\ +\infty & \text{if } n-1 \leq p \leq n+1, \\ \left(1 - \frac{n}{p} \right) \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 + \frac{n}{p-n-1} \left(\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2 \right) + \|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 & \text{if } p \geq n+2. \end{cases} \quad (6)$$

Remark 4.4. While our theoretical insights are inspired by Theorem 1 in Belkin et al. (2020), we extend the framework to classification and OOD data. This extension is non-trivial, as the original result solely focuses on classical supervised regression. Notably, Theorem 1 derives inequalities for the expected OOD risk unlike the equalities found in Belkin et al. (2019) original formulation.

Remark 4.5. We find $\mathbb{E}_{\mathcal{X}}[R_{\text{OOD}}(\hat{f})] = \infty$ around $p = n$, which is characteristic of a double descent phenomenon. This result suggests that OOD scoring functions based on the prediction confidence of binary classifiers $\hat{f}(\cdot)$ exhibit a double descent phenomenon similar to what has been reported for the expected risk (see Appendix A.1). Like the expected risk in the double descent literature (Mei & Montanari, 2022; Louart et al., 2018; Liao et al., 2020; Bach, 2024), this result identifies the ratio p/n as the complexity of a linear model to describe an under- ($p/n < 1$) and an over- ($p/n > 1$) parameterized regimes for the expected OOD risk with a phase transition around $p/n = 1$ characterized by a peak.

5 Empirical Evidence: Double Descent in Practice

In this section, we provide an empirical evaluation of different OOD detection methods with respect to the model width across multiple neural network architectures.

5.1 Toy Example : Gaussian Covariate Model & OOD Detection

We investigate the double descent phenomenon in both generalization and OOD detection using the Gaussian covariate model introduced in Section 4.1. The data is generated from a binary Gaussian Mixture Model (GMM), as illustrated in Figure 1. Our model employs Random ReLU Features (RRF): inputs are mapped to a high-dimensional space via random projections followed by a ReLU activation, and then classified using a linear binary classifier. The OOD samples are drawn from a distinct Gaussian distribution, as described in Section 4.1. By varying the model width, we observe that both the in-distribution Mean-Squared Error (MSE) and the expected OOD risk exhibit double descent behavior, peaking near the interpolation threshold (Figure 1). These results show a concrete empirical manifestation of the theoretical Gaussian setup. Full details are in Appendix D.1.

5.2 Setup of the real case

General Setup. We perform experiments on multiple DNN architectures: ResNet-18 (He et al., 2016), ResNet-34 (Appendix D.4), a 4-block convolutional neural network (CNN), Vision Transformers (ViTs) (Dosovitskiy et al., 2021) and Swin Transformers (Liu et al., 2021b).

Model Setup. To replicate double descent, we follow the experimental setup from Nakkiran et al. (2021), which uses ResNet-18 as the baseline architecture. We apply a similar setup to the 4-block CNN model, ViTs and Swin. We vary the model capacity by altering the hidden dimension (denoted as k) per layer, with values ranging from $k = 1$ to $k = 128$. ResNet-18, which uses a hidden dimension of 64 channels, operates within the overparameterized regime. The depth of the models is kept constant to isolate the effects of width (effective model complexity). The convolutional models are trained using the cross-entropy loss function, with a learning rate of 10^{-4} and the Adam optimizer for 4000 epochs. This extended training regime ensures that models converge for all explored model widths. Moreover, each experiment is conducted five times (with different random seeds). Further details on the experimental setup for the Transformers are given in the Appendix B.2.

Label Noise. To observe the double descent effect, we introduce label noise into the training set by randomly swapping 20% of the labels. This setup simulates real-world scenarios, in which noisy data is common. The models are trained on this noisy dataset but evaluated on a clean test set. Random data augmentations, including random cropping and horizontal flipping, are applied during training. Experiments on the noiseless case are presented in D.6 and OOD risk curves are presented in D.5.

5.3 Evaluation Metrics

We evaluate both generalization and OOD detection using multiple metrics:

- **Generalization:** We report the test accuracy for in-distribution (ID) classification tasks.

- **OOD Detection:** We measure OOD detection performance using the area under the receiver operating characteristic curve (AUC), which is threshold-free and widely adopted in OOD detection research. A higher AUC indicates better performance.
- **Neural Collapse:** We report NC metrics, which, as noted in Ammar et al. (2024); Haas et al. (2023); Zhao & Cao (2023), are associated with certain aspects of OOD detection.

5.4 OOD Datasets

For OOD detection, we evaluate each model using six well-established OOD benchmark datasets: **Textures** (Cimpoi et al., 2014), **Places365** (Zhou et al., 2017), **iNaturalist** (Van Horn et al., 2018), a 10 000 image subset from (Huang & Li, 2021), **ImageNet-O** (Hendrycks et al., 2021) and **SUN** (Xiao et al., 2010). For experiments where CIFAR-10 (or CIFAR-100) is the in-distribution dataset, we also include CIFAR-100 (or CIFAR-10) as an additional OOD benchmark. CIFAR-10/100 contains 50 000 training images and 10 000 test images.

5.5 OOD Detection Methods

In order to have a discussion that generalizes across different OOD detection methods, we evaluate several state-of-the-art methods, categorized by the information they rely on:

- **Logit-based methods:** Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017), Energy scores (Liu et al., 2020), React (Sun et al., 2021), MaxLogit and KL-Matching (Hendrycks et al., 2022),
- **Feature-based methods:** Mahalanobis (Lee et al., 2018b) and Residual (Wang et al., 2022).
- **Hybrid methods:** ViM (Wang et al., 2022), ASH (Djurisic et al., 2023) and NECO (Ammar et al., 2024).

Although the double descent effect is observed in all of our experiments, results from only a few representative methods are presented in the main paper. Additional results can be found in Appendix D.

5.6 OOD Detection and Double Descent

Double Descent & OOD Detection. The primary question addressed in this section is whether the double descent phenomenon extends to OOD detection, as suggested by our theoretical insights. We conduct experiments on CIFAR-10 and CIFAR-100 as ID datasets, and assess OOD detection as a function of model widths. In particular, Figure 2 depicts the evolution of generalization error and OOD detection performance (AUC) for a challenging covariate shift scenario between CIFAR-10 and CIFAR-100. Refer to Appendix D for more results on multiple OOD datasets. Figure 2 shows a double descent phenomenon in all models, with logit-based and hybrid OOD detection methods exhibiting a similar behavior. Experiments evaluating the OOD risk derived from Theorem 1 are provided in Appendix D.5. As shown in the appendix, the OOD risk function also exhibits a double descent phenomenon. This alignment between our theoretical insights and experimental results further reinforces the validity of our analysis.

Feature-Based Techniques & Interpolation Threshold. In some cases, no double descent curve is observed for feature-based techniques. This result suggests that the double descent depends either on the used architecture or the data, as discussed in Appendix E.3. Furthermore, we observe that the interpolation threshold is not always perfectly consistent across OOD datasets or techniques. Those observations are consistent with the Nakkiran et al. (2021)’s results on the CIFAR-10 and CIFAR-100 datasets. Those results suggest that the effective model complexity (EMC) framework (Nakkiran et al., 2021) defined for the generalization error can be extended to OOD detection.

Smaller Models for OOD Detection. Interestingly, in many cases, smaller models are very good OOD detectors. This is particularly relevant for applications with limited computational resources. While techniques such as pruning (Frankle & Carbin, 2019) and quantization (Gholami et al., 2022) can also be used to reduce model size, our study focuses specifically on the impact of model complexity. We believe this behavior may be due to smaller models utilizing their parameters more efficiently. This raises the question of when a lower-complexity model might be more advantageous

than a deep model for OOD detection. The conditions under which this choice becomes optimal will be discussed in Section 5.7.

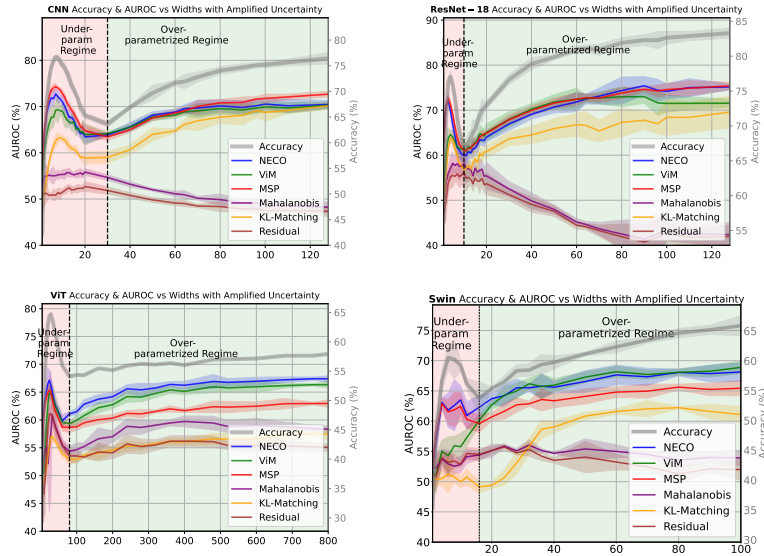


Figure 2: Accuracy and AUC OOD detection metric versus model width. Experiments performed on CNN, ResNet-18, ViT and Swin architectures with CIFAR10 and CIFAR100 as ID and OOD datasets.

Discussion on OOD Methods. OOD detection depends on two key factors: (i) the quality of learned representations, and (ii) the reliability of the confidence score. Logit-based methods rely mainly on model output logits, which are sensitive to model size and complexity (Hendrycks & Gimpel, 2017; Liang et al., 2018), making them closely aligned with the double descent behavior. These methods often show smoother double descent trends. In contrast, feature-based methods depend on how well the model separates ID and OOD data in latent space, which may not always be affected by double descent in the same way.

Discussion on Architectures. Architectural biases significantly affect representation quality and OOD performance (Hein et al., 2019; Tripuraneni et al., 2021a). While all models show double descent, their performance differs: ResNet-18 and Swin improve in the overparameterized regime, CNN performs similarly across regimes, and ViT suffers from poor generalization when overparameterized. To better understand this, we analyze the latent representations using the Neural Collapse framework (Papayan et al., 2020).

5.7 Overparameterization vs. Underparameterization through the Lens of Neural Collapse

From Double Descent to Neural Collapse. We analyze how representation geometry affects the double descent phenomenon in OOD detection. While increased model complexity often improves performance beyond the interpolation threshold, this trend is not universal (Fig. 2). Some models perform better in the underparameterized regime. To explain this, we use the Neural Collapse (NC) framework (Papayan et al., 2020; Ming et al., 2023; Haas et al., 2023; Ammar et al., 2024), which describes how deep models' final-layer features converge to a structured configuration where samples collapse around their respective class centroids.

NC1 Metric for Analyzing Overparameterization. We use the NC1 metric to measure the clustering quality of representations: $NC1 = \text{Tr} \left[\frac{\Sigma_W \Sigma_B^+}{C} \right]$, where Σ_W and Σ_B are the intra- and inter-class covariances, and C is the number of classes. Lower NC1 values indicate better class separation. To quantify the effect of overparameterization, we compute the ratio:

$$NC1_{u/o} = \frac{NC1_u}{NC1_o}, \quad (7)$$

where $NC1_u$ and $NC1_o$ are the NC1 values in the under- and overparameterized regimes optimums, respectively. Higher $NC1_{u/o}$ implies better separation with increased capacity.

Empirical Insights. As shown in Table 1, $NC1_{u/o}$ aligns with OOD performance trends. Overparameterized models with improved $NC1_{u/o} > 1$ typically detect OOD better on the overparametrized regime. Except the CNN architecture, most architectures seem to show improved performance in the overparametrized regime. These findings suggest that representation collapse, as captured by NC1, might indicate OOD detection quality in complex models. Further studies on broader datasets are needed to generalize this insight.

Table 1: Models performance in terms of AUC in the underparametrized local minima (AUC_u) and the overparametrized maximum width (AUC_o), *w.r.t* $NC1_{u/o}$ value. Best is highlighted in **green** when AUC_o is higher, **red** when AUC_u is higher and **blue** if both AUC are within standard deviation range. The highest AUC value per-dataset and per-architecture is highlighted in **bold**.

Model	$NC1_{u/o}$	Method	SUN		Places365		CIFAR-100	
			$AUC_u \uparrow$	$AUC_o \uparrow$	$AUC_u \uparrow$	$AUC_o \uparrow$	$AUC_u \uparrow$	$AUC_o \uparrow$
CNN	0.88	Softmax score	76.09±0.96	75.08±0.75	75.95±0.76	74.59±0.45	74.33±0.32	72.68±0.31
		MaxLogit	72.98±2.22	60.13±0.35	73.33±2.17	61.25±0.42	73.38±0.39	70.37±0.34
		Energy	68.08±3.66	59.73±0.36	69.00±3.56	60.90±0.43	70.78±0.73	70.24±0.35
		Energy+ReAct	59.12±4.73	47.49±0.58	60.79±4.53	49.45±0.52	66.00±1.58	63.57±0.51
		NECO	70.43±2.53	64.22±1.36	71.20±2.59	63.59±0.98	72.40±0.95	70.17±0.76
		ViM	59.94±2.01	59.77±0.66	61.88±1.89	60.93±0.40	69.23±0.78	70.25±0.35
ResNet	1.96	ASH-P	68.60±3.59	60.36±0.37	69.35±3.50	61.48±0.43	71.11±0.53	70.45±0.41
		Softmax score	71.18±0.93	75.82±0.89	71.22±0.93	75.52±0.88	71.21±0.48	75.37±0.42
		MaxLogit	70.64±1.53	72.51±1.03	70.69±1.29	72.64±0.94	69.76±0.39	73.65±0.38
		Energy	69.11±2.49	72.46±1.03	69.19±2.08	72.59±0.94	67.58±0.46	73.61±0.39
		Energy+ReAct	69.57±2.35	71.83±0.88	69.63±1.93	71.97±0.78	67.25±0.91	72.63±0.45
		NECO	70.39±2.30	75.60±1.56	70.46±1.85	75.20±1.42	69.92±0.36	75.28±0.50
Swin	1.70	ViM	66.54±1.99	74.44±0.65	65.38±1.99	73.42±0.65	64.61±0.47	71.54±0.44
		ASH-P	69.11±2.49	71.73 ±1.09	69.19±2.08	71.85±0.98	67.58±0.46	72.89±0.35
		Softmax score	58.82±2.98	67.91±0.69	59.01±2.90	67.66±0.59	61.89±1.42	65.44±0.62
		MaxLogit	59.91±3.11	70.75±0.45	59.84±3.40	70.46±0.46	61.79±1.73	66.95±0.53
		Energy	60.12±3.68	70.79±0.42	58.77±3.98	70.50±0.44	55.52±2.10	66.92±0.51
		Energy+ReAct	60.16±3.79	71.15±0.44	58.85±4.03	70.83±0.48	55.53±2.10	67.27±0.52
ViT	2.32	NECO	64.26±3.20	73.29±0.75	63.88±3.37	72.38±0.66	62.62±2.10	68.13±0.76
		ViM	60.68±2.61	71.69±0.19	58.34±2.60	71.39±0.15	55.95±0.77	68.89±0.51
		ASH-P	59.49±4.21	70.74±0.44	58.15±4.42	70.42±0.40	55.10±2.28	66.89±0.55
		Softmax score	66.28±0.19	64.87±0.27	66.26±0.36	64.61±0.26	65.18±0.38	62.96±0.33
		MaxLogit	66.09±1.48	70.30±0.46	66.13±1.50	69.79±0.26	64.60±0.35	66.69±0.39
		Energy	64.79±2.81	70.50±0.48	64.86±2.65	69.98±0.26	63.08±0.44	66.79±0.38
ViT	2.32	Energy+ReAct	64.51±2.93	70.51±0.49	64.65±2.75	69.97±0.26	62.86±0.58	66.78±0.39
		NECO	67.61±1.61	75.89±0.47	67.47±1.68	74.29±0.29	66.28±0.54	67.40±0.27
		ViM	63.14±3.54	72.25±0.37	63.30±3.36	71.41±0.15	64.81±0.65	66.34±0.30
		ASH-P	64.79±2.81	70.27±0.50	64.86±2.65	69.79±0.25	63.08±0.44	66.61±0.36

Empirical Evidence of Depth-Wise Double Descent. We also investigate the influence of model depth and assess OOD detection and accuracy as functions of model depth with fixed width in Appendix F. Our results highlight similar double descent patterns in both ID generalization and post-hoc OOD detection. To the best of our knowledge, these are the first results illustrating a depth-wise double descent.

6 Conclusion

In this paper, we presented empirical and theoretical insight that the double descent phenomenon also impacts OOD detection across architectures. Using Random Matrix Theory we showed that both prediction and OOD risks peak at the interpolation threshold for least-squares classifiers on Gaussian data. We have studied the double descent phenomenon across various architectures and OOD detection algorithms. Additionally, we provided insights into why the overparameterized regime can perform better or worse than the underparameterized one. These observations emphasize the critical role of latent representations in OOD detection under different levels of model complexity. While our theoretical analysis relies on simplified settings (e.g., linear models and Gaussian inputs), future works may investigate more complex systems and include additional factors such as training dynamics, or regularization.

Limitations: Finally, our study is limited to post-hoc out-of-distribution (OOD) detection; exploring alternative OOD detection paradigms could be a valuable direction for future research.

References

- Ammar, M. B., Belkhir, N., Popescu, S., Manzanera, A., and Franchi, G. NECO: NEural Collapse Based Out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=9R0uKblmi7>.
- Bach, F. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024. doi: 10.1137/23M1558781. URL <https://doi.org/10.1137/23M1558781>.
- Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, April 2020. ISSN 1091-6490. doi: 10.1073/pnas.1907378117. URL <http://dx.doi.org/10.1073/pnas.1907378117>.
- Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 541–549. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/belkin18a.html>.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, July 2019. ISSN 1091-6490. doi: 10.1073/pnas.1903070116. URL <http://dx.doi.org/10.1073/pnas.1903070116>.
- Belkin, M., Hsu, D., and Xu, J. Two Models of Double Descent for Weak Features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, January 2020. ISSN 2577-0187. doi: 10.1137/20m1336072. URL <http://dx.doi.org/10.1137/20m1336072>.
- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. URL <https://doi.org/10.1007/s10994-009-5152-4>.
- Bendale, A. and Boulton, T. E. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- Bengio, Y. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009. ISSN 1935-8237. doi: 10.1561/22000000006. URL <https://doi.org/10.1561/22000000006>.
- Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Occam’s razor. *Information Processing Letters*, 24(6):377–380, 1987. ISSN 0020-0190. doi: [https://doi.org/10.1016/0020-0190\(87\)90114-1](https://doi.org/10.1016/0020-0190(87)90114-1). URL <https://www.sciencedirect.com/science/article/pii/0020019087901141>.
- Breiman, L. and Freedman, D. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983. doi: 10.1080/01621459.1983.10477941. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10477941>.
- Brellmann, D., Berthier, E., Filliat, D., and Frehse, G. On double descent in reinforcement learning with LSTD and random features. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=9RIbNm984>.
- Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1): 2914, 2021. URL <https://doi.org/10.1038/s41467-021-23103-1>.
- Cherkassky, V. and Lee, E. H. To understand double descent, we need to understand vc theory. *Neural Networks*, 169:242–256, 2024. ISSN 0893-6080. doi: <https://doi.org/10.1016/j.neunet.2023.10.014>. URL <https://www.sciencedirect.com/science/article/pii/S0893608023005658>.
- Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

- Couillet, R. and Liao, Z. *Random matrix methods for machine learning*. Cambridge University Press, 2022.
- D’Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance(s) in the lazy regime. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2280–2290. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/d-ascoli20a.html>.
- Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.
- Dereziński, M., Liang, F., and Mahoney, M. W. Exact expressions for double descent and implicit regularization via surrogate random design. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- DeVries, T. and Taylor, G. W. Learning confidence for out-of-distribution detection in neural networks, 2018. URL <https://arxiv.org/abs/1802.04865>.
- Djurisic, A., Bozanic, N., Ashok, A., and Liu, R. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=ndYXTEL6cZz>.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- Drummond, N. and Shearer, R. The open world assumption. In *eSI workshop: the closed world of databases meets the open world of the semantic web*, volume 15, pp. 1, 2006.
- Fang, Z., Li, Y., Liu, F., Han, B., and Lu, J. On the learnability of out-of-distribution detection. *Journal of Machine Learning Research*, 25(84):1–83, 2024. URL <http://jmlr.org/papers/v25/23-1257.html>.
- Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rJl-b3RcF7>.
- Geman, S., Bienenstock, E., and Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4:1–58, 01 1992. doi: 10.1162/neco.1992.4.1.1.
- Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pp. 291–326. Chapman and Hall/CRC, 2022.
- Haas, J., Yolland, W., and Rabus, B. T. Linking Neural Collapse and L2 Normalization with Improved Out-of-Distribution Detection in Deep Neural Networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=fjkN5Ur2d6>.
- Hao, Y. and Zhang, T. The Surprising Harmfulness of Benign Overfitting for Adversarial Robustness, 2024. URL <https://arxiv.org/abs/2401.12236>.
- Hao, Y., Lin, Y., Zou, D., and Zhang, T. On the Benefits of Over-parameterization for Out-of-Distribution Generalization, 2024. URL <https://arxiv.org/abs/2403.17592>.
- Hastie, T. *The elements of statistical learning: data mining, inference, and prediction*, 2009.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

- Hein, M., Andriushchenko, M., and Bitterwolf, J. 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*, pp. 41–50, 2019.
- Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Hkg4TI9xl>.
- Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021.
- Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. X. Scaling Out-of-Distribution Detection for Real-World Settings. In *International Conference on Machine Learning*, 2022. URL <https://api.semanticscholar.org/CorpusID:227407829>.
- Huang, R. and Li, Y. MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8710–8719, June 2021.
- Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4631–4640. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/jacot20a.html>.
- Kausik, C., Srivastava, K., and Sonthalia, R. Double descent and overfitting under noisy inputs and distribution shift for linear denoisers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=HxfqTdLIRF>.
- Kini, G. R. and Thrampoulidis, C. Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2527–2532. IEEE, 2020.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015. URL <https://doi.org/10.1038/nature14539>.
- Lee, E. H. and Cherkassky, V. VC Theoretical Explanation of Double Descent, 2022. URL <https://arxiv.org/abs/2205.15549>.
- Lee, K., Lee, K., Lee, H., and Shin, J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and adversarial attacks. In *NeurIPS*, 2018a.
- Lee, K., Lee, K., Lee, H., and Shin, J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf.
- Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=H1VGkIxRZ>.

- Liao, Z., Couillet, R., and Mahoney, M. W. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13939–13950. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf.
- Liu, F., Liao, Z., and Suykens, J. Kernel regression in high dimensions: Refined analysis beyond double descent. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 649–657. PMLR, 13–15 Apr 2021a. URL <https://proceedings.mlr.press/v130/liu21b.html>.
- Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.
- Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. ISSN 10505164, 21688737. URL <https://www.jstor.org/stable/26542333>.
- Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: <https://doi.org/10.1002/cpa.22008>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22008>.
- Ming, Y., Sun, Y., Dia, O., and Li, Y. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=aEFaE0W5pAd>.
- Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, dec 2021. doi: 10.1088/1742-5468/ac3a74. URL <https://doi.org/10.1088/1742-5468/ac3a74>.
- Olmin, A. and Lindsten, F. Towards understanding epoch-wise double descent in two-layer linear neural networks, 2024. URL <https://arxiv.org/abs/2407.09845>.
- Papayan, V., Han, X. Y., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, September 2020. ISSN 1091-6490. doi: 10.1073/pnas.2015509117. URL <http://dx.doi.org/10.1073/pnas.2015509117>.
- Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., and Boulton, T. E. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.
- Stephenson, C. and Lee, T. When and how epochwise double descent happens, 2021. URL <https://arxiv.org/abs/2108.12006>.
- Sun, Y., Guo, C., and Li, Y. ReAct: Out-of-distribution Detection With Rectified Activations. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=IBVBtz_sRSm.

- Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/sun22d.html>.
- Tripuraneni, N., Adlam, B., and Pennington, J. Overparameterization improves robustness to covariate shift in high dimensions. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13883–13897. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/73fed7fd472e502d8908794430511f4d-Paper.pdf.
- Tripuraneni, N., Adlam, B., and Pennington, J. Overparameterization improves robustness to covariate shift in high dimensions. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13883–13897. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/73fed7fd472e502d8908794430511f4d-Paper.pdf.
- Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.
- Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930, June 2022.
- Wang, M. and Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2018.05.083>. URL <https://www.sciencedirect.com/science/article/pii/S0925231218306684>.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. URL <https://api.semanticscholar.org/CorpusID:1309931>.
- Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. URL <https://doi.org/10.1007/s11263-024-02117-4>.
- Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10767–10777. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/yang20j.html>.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, February 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL <https://doi.org/10.1145/3446776>.
- Zhang, J., Chen, Y., Jin, C., Zhu, L., and Gu, Y. Epa: Neural collapse inspired robust out-of-distribution detector. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6515–6519. IEEE, 2024.
- Zhao, Z. and Cao, L. Dual representation learning for out-of-distribution detection. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=PHAr3q49h6>.
- Zhou, B., Lapedriza, A., Torralba, A., and Oliva, A. Places: An Image Database for Deep Scene Understanding. *Journal of Vision*, 17(10):296–296, 2017.

Double Descent Meets Out-of-Distribution Detection: Theoretical Insights and Empirical Analysis on the Role of Model Complexity

Table of Contents - Appendix

A Proof of theorem	16
A.1 Proof of Theorem 2	17
A.2 Proof of Theorem 1	17
B Details about Baselines and Architectures Implementations	22
B.1 Baselines OOD Methods	22
B.2 Experiments Setup	23
C Details about Neural Collapse	23
D Complementary Results on Double Descent and OOD Detection	24
D.1 ReLU Random Feature Model Experimental Setting	24
D.2 CIFAR-10 Additional Results	24
D.3 CIFAR-100 Results	25
D.4 ResNet-34 Results	25
D.5 Out-of-Distribution Risk Experiments	26
D.6 Experiments Without Label-Noise Results	26
E Model Representations and Neural Collapse Analysis	26
E.1 Structure of the Model Representation	26
E.2 Evolution of NC1	26
E.3 Mahalanobis and Residual Joint Performance	27
F Depth-wise Double Descent	27
F.1 Experimental Setup	28
F.2 Results	28
F.3 Connection to Width-wise Experiments.	28

A Proof of Theorems

While our theoretical insights are inspired by Theorem 1 in Belkin et al. (2020), we extend the framework to classification and OOD data.

A.1 Proof for the Expected Risk

Leveraging the Random Matrix Theory and extending the Theorem 1 in Belkin et al. (2020), we derive bounds in Theorem 2 for the expected risk of the subset of classifiers defined in $\hat{\mathcal{F}}_d$ with equation 4. This section is dedicated to the proof of Theorem 2 and follows the same line of arguments of Theorem 1 in Belkin et al. (2020).

Theorem 2. Let $(p, q) \in [d]^2$ such that $p + q = d$, $\mathcal{T} \subseteq [d]$ with $|\mathcal{T}| = p$ an arbitrary subset of the d first natural integers, and $\mathcal{T}^c := [d] \setminus \mathcal{T}$ its complement set. Let $\hat{\mathbf{w}} \in \mathbb{R}^d(\mathcal{T})$ such that $\hat{\mathbf{w}}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^+ \mathbf{z} \in \mathbb{R}^p$ and $\hat{\mathbf{w}}_{\mathcal{T}^c} = \mathbf{0}_q \in \mathbb{R}^q$. Under Assumptions 4.1-4.2, the expected risk with respect to the loss function $\ell : (\hat{y}, y) \mapsto (\hat{y} - y)^2$ of the predictor $\hat{f}_{\mathcal{T}} : \mathbf{x} \mapsto \phi(\mathbf{x}^T \hat{\mathbf{w}})$ satisfies

$$c(n, p) \leq \mathbb{E}_{\mathbf{X}} [R(\hat{f}_{\mathcal{T}})] \leq C c(n, p),$$

where $c, C > 0$ and

$$c(n, p) = \begin{cases} \frac{p}{n-p-1} \left(\|\mathbf{w}_{\mathcal{T}^c}^*\|_2^2 + \sigma^2 \right) + \|\mathbf{w}_{\mathcal{T}^c}^*\|_2^2 & \text{if } p \leq n-2, \\ +\infty & \text{if } n-1 \leq p \leq n+1, \\ \left(1 - \frac{n}{p}\right) \|\mathbf{w}_{\mathcal{T}}^*\|_2^2 + \frac{n}{p-n-1} \left(\|\mathbf{w}_{\mathcal{T}^c}^*\|_2^2 + \sigma^2 \right) + \|\mathbf{w}_{\mathcal{T}^c}^*\|_2^2 & \text{if } p \geq n+2. \end{cases} \quad (8)$$

Remark A.1. Theorem 1 in Belkin et al. (2020) constitutes a special case of Theorem 2 for $\phi : \mathbf{x} \mapsto \mathbf{x}$, which corresponds to the case where $\Sigma = \mathbf{I}_d$.

Remark A.2. From Theorem 2, we have $\mathbb{E}_{\mathbf{X}} [R(\hat{f}_{\mathcal{T}})] = \infty$ around $p = n$. The expected risk decreases again as p increases beyond n and highlights a double descent phenomenon. This result is consistent with the literature of double descent (Mei & Montanari, 2022; Louart et al., 2018; Liao et al., 2020; Bach, 2024), which identifies the ratio p/n as the complexity of a linear model to describe an under- ($p/n < 1$) and an over- ($p/n > 1$) parameterized regimes for the expected risk with a phase transition around $p/n = 1$ characterized by a peak.

Proof. Let $\mathbf{x} \in \mathcal{X}$ and

$$\Delta_{\mathbf{w}} = \hat{\mathbf{w}} - \mathbf{w}^*.$$

We have

$$\hat{f}_{\mathcal{T}}(\mathbf{x}) = \phi(\mathbf{x}^T \hat{\mathbf{w}}) = \phi(\mathbf{x}^T \mathbf{w}^*) + \mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*)).$$

From the mean-value theorem, there exists

$$c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*) \in (\min(\mathbf{x}^T \hat{\mathbf{w}}, \mathbf{x}^T \mathbf{w}^*), \max(\mathbf{x}^T \hat{\mathbf{w}}, \mathbf{x}^T \mathbf{w}^*)),$$

such that

$$\phi(\mathbf{x}^T (\mathbf{w}^* + \Delta_{\mathbf{w}})) = \phi(\mathbf{x}^T \mathbf{w}^*) + \mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*)).$$

We have thus

$$\begin{aligned} \mathbb{E}_{\mathbf{X}} [R(\hat{f}_{\mathcal{T}})] &= \mathbb{E}_{\mathbf{x}, \mathbf{X}} \left[(\hat{f}(\mathbf{x}) - f^*(\mathbf{x}))^2 \right] = \mathbb{E}_{\mathbf{x}, \mathbf{X}} \left[(\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^*))^2 \right] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{X}} \left[(\mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*)))^2 \right] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{X}} \left[\Delta_{\mathbf{w}}^T \mathbf{x} \mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*))^2 \right] \\ &= \text{Tr} \left(\mathbb{E}_{\mathbf{x}, \mathbf{X}} [\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*))^2 \mathbf{x} \mathbf{x}^T \Delta_{\mathbf{w}} \Delta_{\mathbf{w}}^T] \right) \\ &= \text{Tr} \left(\mathbb{E}_{\mathbf{X}} [\mathbb{E}_{\mathbf{x}} [\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*))^2 \mathbf{x} \mathbf{x}^T] \Delta_{\mathbf{w}} \Delta_{\mathbf{w}}^T] \right) \\ &= \text{Tr} \left(\mathbb{E}_{\mathbf{X}} [\Sigma \Delta_{\mathbf{w}} \Delta_{\mathbf{w}}^T] \right) \\ &= \mathbb{E}_{\mathbf{X}} [\Delta_{\mathbf{w}}^T \Sigma \Delta_{\mathbf{w}}], \end{aligned}$$

where

$$\begin{aligned}\Sigma &= \mathbb{E}_{\mathbf{x}} [\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^*))^2 \mathbf{x} \mathbf{x}^T] \\ &= \mathbb{E}_{\mathbf{x}} \left[\left(\frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^*)}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^*} \right)^2 \mathbf{x} \mathbf{x}^T \right].\end{aligned}$$

From the min-max theorem, we have

$$\mathbb{E}_{\mathbf{X}} [\lambda_{\min}(\Sigma) \Delta_w^T \Delta_w] \leq \mathbb{E}_{\mathbf{X}} [\Delta_w^T \Sigma \Delta_w] \leq \mathbb{E}_{\mathbf{X}} [\lambda_{\max}(\Sigma) \Delta_w^T \Delta_w].$$

From Lemma 1.1, there exists $c, C > 0$ independent of $\hat{\mathbf{w}}$ (and thus $\mathbf{X}$) such that

$$c \mathbb{E}_{\mathbf{X}} [\Delta_w^T \Delta_w] \leq \mathbb{E}_{\mathbf{X}} [\Delta_w^T \Sigma \Delta_w] \leq C \mathbb{E}_{\mathbf{X}} [\Delta_w^T \Delta_w].$$

For $\mathbb{E}_{\mathbf{X}} [\Delta_w^T \Delta_w]$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{X}} [\Delta_w^T \Delta_w] &= \mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2] \\ &= \mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^*\|_2^2] + \mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}^c} - \mathbf{w}_{\mathcal{T}^c}^*\|_2^2] \\ &= \mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^*\|_2^2] + \|\mathbf{w}_{\mathcal{T}^c}^*\|_2^2.\end{aligned}$$

From Lemma 1.2, we have

$$\mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^*\|_2^2] = \begin{cases} \frac{p}{n-p-1} (\|\mathbf{w}_{\mathcal{T}^c}^*\|_2^2 + \sigma^2) & \text{if } p \leq n-2, \\ +\infty & \text{if } n-1 \leq p \leq n+1, \\ (1 - \frac{n}{p}) \|\mathbf{w}_{\mathcal{T}}^*\|_2^2 + \frac{n}{p-n-1} (\|\mathbf{w}_{\mathcal{T}^c}^*\|_2^2 + \sigma^2) & \text{if } p \geq n+2, \end{cases}$$

which concludes the proof. $\square$

A.2 Proof for the Expected OOD Risk (Theorem 1)

This section is dedicated to the proof of Theorem 1.

Theorem 1. Let $(p, q) \in [d]^2$ such that $p + q = d$, $\mathcal{T} \subseteq [d]$ with $|\mathcal{T}| = p$ an arbitrary subset of the d first natural integers, and $\mathcal{T}^c := [d] \setminus \mathcal{T}$ its complement set. Let $\hat{\mathbf{w}} \in R^d(\mathcal{T})$ such that $\hat{\mathbf{w}}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^+ \mathbf{z} \in \mathbb{R}^p$ and $\hat{\mathbf{w}}_{\mathcal{T}^c} = \mathbf{0}_q \in \mathbb{R}^q$. Under Assumptions 4.1-4.2, the expected OOD risk on the predictor $\hat{f}_{\mathcal{T}} : \mathbf{x} \mapsto \phi(\mathbf{x}^T \hat{\mathbf{w}})$ satisfies

$$c \, c(n, p) \leq \mathbb{E}_{\mathbf{X}} [R_{\text{OOD}}(\hat{f})] \leq C \, c(n, p),$$

where $c, C > 0$ and

$$c(n, p) = \begin{cases} \frac{p}{n-p-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2) + \|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 & \text{if } p \leq n-2, \\ +\infty & \text{if } n-1 \leq p \leq n+1, \\ (1 - \frac{n}{p}) \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 + \frac{n}{p-n-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2) + \|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 & \text{if } p \geq n+2. \end{cases}$$

Proof. The layout of the proof is similar to the proof of Theorem 2. Let $\mathbf{x} \in \mathcal{X}$ and

$$\Delta_w = \hat{\mathbf{w}} - \mathbf{w}^{\text{OOD}}.$$

From the mean-value theorem, there exists

$$c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}) \in (\min(\mathbf{x}^T \hat{\mathbf{w}}, \mathbf{x}^T \mathbf{w}^{\text{OOD}}), \max(\mathbf{x}^T \hat{\mathbf{w}}, \mathbf{x}^T \mathbf{w}^{\text{OOD}})),$$

such that

$$\begin{aligned}\hat{f}_{\mathcal{T}}(\mathbf{x}) &= \phi(\mathbf{x}^T \hat{\mathbf{w}}) = \phi(\mathbf{x}^T (\mathbf{w}^{\text{OOD}} + \Delta_w)) \\ &= \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}}) + \mathbf{x}^T \Delta_w \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}})).\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[R_{\text{OOD}}(\hat{f})] &= \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}, \mathbf{X}}[(\hat{f}(\mathbf{x}) - f^{\text{OOD}}(\mathbf{x}))^2] + \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}, \mathbf{X}}[(\hat{f}(\mathbf{x}) - f^{\text{OOD}}(\mathbf{x}))^2] \\
&= \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}, \mathbf{X}}[(\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}}))^2] + \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}, \mathbf{X}}[(\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}}))^2] \\
&= \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}, \mathbf{X}}[(\mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}})))^2] + \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}, \mathbf{X}}[(\mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}})))^2] \\
&= \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}, \mathbf{X}}[\Delta_{\mathbf{w}}^T \mathbf{x} \mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2] + \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}, \mathbf{X}}[\Delta_{\mathbf{w}}^T \mathbf{x} \mathbf{x}^T \Delta_{\mathbf{w}} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2] \\
&= \text{Tr}(\mathbb{E}_{\mathbf{X}}[\mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}}[\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 \mathbf{x} \mathbf{x}^T] + \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}}[\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 \mathbf{x} \mathbf{x}^T]] \Delta_{\mathbf{w}} \Delta_{\mathbf{w}}^T) \\
&= \text{Tr}(\mathbb{E}_{\mathbf{X}}[\Sigma + \Sigma^{\text{OOD}}] \Delta_{\mathbf{w}} \Delta_{\mathbf{w}}^T) \\
&= \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma \Delta_{\mathbf{w}}] + \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma^{\text{OOD}} \Delta_{\mathbf{w}}],
\end{aligned}$$

where

$$\begin{aligned}
\Sigma &= \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}}[\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 \mathbf{x} \mathbf{x}^T] \\
&= \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}}[\left(\frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}})}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^{\text{OOD}}}\right)^2 \mathbf{x} \mathbf{x}^T]
\end{aligned}$$

and

$$\begin{aligned}
\Sigma^{\text{OOD}} &= \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}}[\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 \mathbf{x} \mathbf{x}^T] \\
&= \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}}[\left(\frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}})}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^{\text{OOD}}}\right)^2 \mathbf{x} \mathbf{x}^T].
\end{aligned}$$

From the min-max theorem, we have

$$\mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma \Delta_{\mathbf{w}}] + \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma^{\text{OOD}} \Delta_{\mathbf{w}}] \geq \mathbb{E}_{\mathbf{X}}[(\lambda_{\min}(\Sigma) + \lambda_{\min}(\Sigma^{\text{OOD}})) \Delta_{\mathbf{w}}^T \Delta_{\mathbf{w}}]$$

and

$$\mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma \Delta_{\mathbf{w}}] + \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma^{\text{OOD}} \Delta_{\mathbf{w}}] \leq \mathbb{E}_{\mathbf{X}}[(\lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma^{\text{OOD}})) \Delta_{\mathbf{w}}^T \Delta_{\mathbf{w}}].$$

From Lemma 1.1, there exists $c, C > 0$ independent of $\hat{\mathbf{w}}$ (and thus $\mathbf{X}$) such that

$$c \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Delta_{\mathbf{w}}] \leq \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma \Delta_{\mathbf{w}}] + \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Sigma^{\text{OOD}} \Delta_{\mathbf{w}}] \leq C \mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Delta_{\mathbf{w}}].$$

For $\mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Delta_{\mathbf{w}}]$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\Delta_{\mathbf{w}}^T \Delta_{\mathbf{w}}] &= \mathbb{E}_{\mathbf{X}}[\|\hat{\mathbf{w}} - \mathbf{w}^{\text{OOD}}\|_2^2] \\
&= \mathbb{E}_{\mathbf{X}}[\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] + \mathbb{E}_{\mathbf{X}}[\|\hat{\mathbf{w}}_{\mathcal{T}^c} - \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2] \\
&= \mathbb{E}_{\mathbf{X}}[\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] + \|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2.
\end{aligned}$$

From Lemma 1.2, we have

$$\mathbb{E}_{\mathbf{X}}[\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] = \begin{cases} \frac{p}{n-p-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2) & \text{if } p \leq n-2, \\ +\infty & \text{if } n-1 \leq p \leq n+1, \\ (1 - \frac{n}{p}) \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 + \frac{n}{p-n-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2) & \text{if } p \geq n+2, \end{cases}$$

which concludes the proof. $\square$

Lemma 1.1. Let $\Sigma, \Sigma^{\text{OOD}} \in \mathbb{R}^{d \times d}$ defined as

$$\Sigma = \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}}[\left(\frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}})}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^{\text{OOD}}}\right)^2 \mathbf{x} \mathbf{x}^T]$$

and

$$\Sigma^{\text{OOD}} = \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}, \mathcal{Y}}^{\text{OOD}}} \left[\left(\frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}})}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^{\text{OOD}}} \right)^2 \mathbf{x} \mathbf{x}^T \right].$$

Under Assumption 4.1-4.2, the matrices Σ and Σ^{OOD} are positive-definite. Furthermore, there exists $c_1, C_1 > 0$ and $c_2, C_2 > 0$ independent of $\hat{\mathbf{w}}$ such that

$$\lambda_{\min}(\Sigma) \geq c_1 \quad \text{and} \quad \lambda_{\max}(\Sigma) \leq C_1$$

and

$$\lambda_{\min}(\Sigma^{\text{OOD}}) \geq c_2 \quad \text{and} \quad \lambda_{\max}(\Sigma^{\text{OOD}}) \leq C_2.$$

Proof. From the mean-value theorem, there exists

$$c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}) \in (\min(\mathbf{x}^T \hat{\mathbf{w}}, \mathbf{x}^T \mathbf{w}^{\text{OOD}}), \max(\mathbf{x}^T \hat{\mathbf{w}}, \mathbf{x}^T \mathbf{w}^{\text{OOD}})),$$

such that

$$\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}})) = \frac{\phi(\mathbf{x}^T \hat{\mathbf{w}}) - \phi(\mathbf{x}^T \mathbf{w}^{\text{OOD}})}{\mathbf{x}^T \hat{\mathbf{w}} - \mathbf{x}^T \mathbf{w}^{\text{OOD}}}.$$

Using $c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}})$, we rewrite the matrix Σ as

$$\Sigma = \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 \mathbf{x} \mathbf{x}^T].$$

The matrix Σ is semi-positive-definite. Let $\mu(\cdot)$ denotes the probability density function of $P_{\mathcal{X}, \mathcal{Y}}$.

Let $\mathbf{u} = \arg \min_{\mathbf{a} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}} \frac{\mathbf{a}^T \Sigma \mathbf{a}}{\mathbf{a}^T \mathbf{a}}$. From the min-max theorem, we have

$$\lambda_{\min}(\Sigma) = \frac{\mathbf{u}^T \Sigma \mathbf{u}}{\mathbf{u}^T \mathbf{u}}.$$

Since the derivative of $\phi(\cdot)$ is strictly positive, we have

$$\begin{aligned} \lambda_{\min}(\Sigma) &= \frac{\mathbf{u}^T \Sigma \mathbf{u}}{\mathbf{u}^T \mathbf{u}} = \frac{1}{\mathbf{u}^T \mathbf{u}} \mathbf{u}^T \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 \mathbf{x} \mathbf{x}^T] \mathbf{u} \\ &= \frac{1}{\mathbf{u}^T \mathbf{u}} \int_{\mathbb{R}^d} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 (\mathbf{u}^T \mathbf{x})^2 \mu(\mathbf{x}) d\mathbf{x} \\ &\geq \underbrace{\frac{1}{\mathbf{u}^T \mathbf{u}} \int_{R^d(\mathcal{T}^c)} \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 (\mathbf{u}^T \mathbf{x})^2 \mu(\mathbf{x}) d\mathbf{x}}_{=c}. \end{aligned}$$

Note that for all $\mathbf{x} \in R^d(\mathcal{T}^c)$, we have

$$\hat{\mathbf{w}}^T \mathbf{x} = \hat{\mathbf{w}}_{\mathcal{T}}^T \mathbf{x}_{\mathcal{T}} + \hat{\mathbf{w}}_{\mathcal{T}^c}^T \mathbf{x}_{\mathcal{T}^c} = \hat{\mathbf{w}}_{\mathcal{T}}^T \mathbf{0}_p + \mathbf{0}_{d-p}^T \mathbf{x}_{\mathcal{T}^c} = 0.$$

We deduce that c is independent of $\hat{\mathbf{w}}$ since

$$c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}) \in (\min(0, \mathbf{x}^T \mathbf{w}^{\text{OOD}}), \max(0, \mathbf{x}^T \mathbf{w}^{\text{OOD}})), \quad \forall \mathbf{x} \in R^d(\mathcal{T}^c).$$

Furthermore, $c > 0$ since $\mathbf{x} \mapsto \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 (\mathbf{u}^T \mathbf{x})^2$ is nonnegative and is not the zero map on $R^d(\mathcal{T})$ as $\mathbf{u} \notin R^d(\mathcal{T})$ (Assumption 4.2). We conclude there exists $c > 0$ independent of $\hat{\mathbf{w}}$ such that $\lambda_{\min}(\Sigma) \geq c$.

Next, we are going to show that there exists $C > 0$ independent of $\hat{\mathbf{w}}$ such that $\lambda_{\max}(\Sigma) \leq C$. From Assumption 4.1, we know there exists $K' > 0$ such that for all $x \in \mathbb{R}$ we have

$$\phi'(x) \leq K'.$$

Let $K > K'$ and $\mathbf{M} = K^2 \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [\mathbf{x} \mathbf{x}^T] - \Sigma$. $\mathbf{M}$ is a symmetric matrix. From the min-max theorem $\mathbf{M}$ is semi positive-definite if for all $\mathbf{a} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$, we have

$$\mathbf{a}^T \mathbf{M} \mathbf{a} \geq 0.$$

Let $\mathbf{a} \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$, we have

$$\begin{aligned} \mathbf{a}^T \mathbf{M} \mathbf{a} &= \mathbf{a}^T [K^2 \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [\mathbf{x} \mathbf{x}^T] - \Sigma] \mathbf{a} \\ &= \mathbf{a}^T [\mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [K^2 \mathbf{x} \mathbf{x}^T] - \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [\phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2 \mathbf{x} \mathbf{x}^T]] \mathbf{a} \\ &= \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} \left[\underbrace{(K^2 - \phi'(c(\mathbf{x}, \hat{\mathbf{w}}, \mathbf{w}^{\text{OOD}}))^2)}_{>0} \underbrace{(\mathbf{a}^T \mathbf{x})^2}_{\geq 0} \right] \\ &\geq 0. \end{aligned}$$

We deduce that $\mathbf{M} \succeq 0$. Therefore, $\Sigma \preceq K^2 \mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [\mathbf{x} \mathbf{x}^T]$ and we have

$$\lambda_{\max}(\Sigma) \leq \underbrace{K^2 \lambda_{\max}(\mathbb{E}_{(\mathbf{x}, \cdot) \sim P_{\mathcal{X}, \mathcal{Y}}} [\mathbf{x} \mathbf{x}^T])}_{=C}.$$

C is independent of $\hat{\mathbf{w}}$. With a similar proof, we can derive a result for Σ^{OOD} . $\square$

Lemma 1.2. Let $(p, q) \in [d]^2$ such that $p + q = d$, $\mathcal{T} \subseteq [d]$ an arbitrary subset of the d first natural integers, and $\mathcal{T}^c := [d] \setminus \mathcal{T}$ its complement set. Let $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that $\hat{\mathbf{w}}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{z} \in \mathbb{R}^p$ and $\hat{\mathbf{w}}_{\mathcal{T}^c} = \mathbf{0}_q \in \mathbb{R}^q$. We have

$$\mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] = \begin{cases} \frac{p}{n-p-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2) & \text{if } p \leq n-2, \\ +\infty & \text{if } n-1 \leq p \leq n+1, \\ (1 - \frac{n}{p}) \|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \frac{n}{p-n-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2) & \text{if } p \geq n+2. \end{cases}$$

Proof. Let $\boldsymbol{\eta} = \mathbf{z} - \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}$. Since $\hat{\mathbf{w}}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{z}$, we have $\hat{\mathbf{w}}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^{\perp} (\boldsymbol{\eta} + \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}})$. Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbf{X}} [\|\mathbf{w}_{\mathcal{T}}^{\text{OOD}} - \hat{\mathbf{w}}_{\mathcal{T}}\|_2^2] &= \mathbb{E}_{\mathbf{X}} [\|(I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}} - \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2] \\ &= \mathbb{E}_{\mathbf{X}} [\|(I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 + \|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2 - 2 \langle (I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}, \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} \rangle]. \end{aligned}$$

Since $(\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}})^T = \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}$ and $(\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}})^T \mathbf{X}_{\mathcal{T}}^{\perp} = \mathbf{X}_{\mathcal{T}}^{\perp}$, we have

$$\begin{aligned} \langle (I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}, \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} \rangle_2 &= ((I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}})^T \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} \\ &= [\mathbf{w}_{\mathcal{T}}^{\text{OOD}}]^T (I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}})^T \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} \\ &= [\mathbf{w}_{\mathcal{T}}^{\text{OOD}}]^T \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} - [\mathbf{w}_{\mathcal{T}}^{\text{OOD}}]^T (\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} \\ &= [\mathbf{w}_{\mathcal{T}}^{\text{OOD}}]^T \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} - [\mathbf{w}_{\mathcal{T}}^{\text{OOD}}]^T \mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta} \\ &= 0. \end{aligned}$$

$(I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}$ and $\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}$ are thus orthogonal. We deduce that

$$\begin{aligned} \mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] &= \mathbb{E}_{\mathbf{X}} [\|(I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 + \|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2] \\ &= \mathbb{E}_{\mathbf{X}} [\|(I_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] + \mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2]. \end{aligned} \quad (9)$$

Leveraging the same arguments used by Belkin et al. (2020) to prove the existence of the double descent phenomenon in the regression problem, we distinguish two cases depending on n and p to derive equation 9.

Classical Regime ($p < n$). Breiman & Freedman (1983) studied this regime for the regression problem and found:

$$\mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] = \frac{p}{n-p-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2).$$

Interpolating Regime ($p \geq n$). The interpolating regime has been considered in Belkin et al. (2020) for the regression problem. We can observe that:

$$\mathbf{w}_{\mathcal{T}}^{\text{OOD}} = (\mathbf{I}_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}} + \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}$$

and

$$\begin{aligned} \langle (\mathbf{I}_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}, \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}} \rangle &= (\mathbf{I}_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}})^T (\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}) \\ &= [\mathbf{w}_{\mathcal{T}}^{\text{OOD}}]^T (\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} - (\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}})(\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}})) \mathbf{w}_{\mathcal{T}}^{\text{OOD}} \\ &= 0. \end{aligned}$$

Since $(\mathbf{I}_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}$ and $\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}$ are orthogonal, we have

$$\|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 = \|(\mathbf{I}_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 + \|\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 \quad (\text{Pythagorean theorem})$$

and thus

$$\|(\mathbf{I}_p - \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}}) \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 = \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 - \|\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2.$$

Putting the equation above into equation 9, we obtain

$$\mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] = \mathbb{E}_{\mathbf{X}} [\|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] - \mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] + \mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2]. \quad (10)$$

Note that $\mathbf{\Pi}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}}^T (\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T)^{-1} \mathbf{X}_{\mathcal{T}}$ is the orthogonal projection matrix for the row space of $\mathbf{X}_{\mathcal{T}}$. We can thus write $\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}} = \mathbf{\Pi}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}$ as a linear combination of rows of $\mathbf{X}_{\mathcal{T}}$. Then, using the fact that the x_i in $\mathbf{X}$ are i.i.d. and drawn from a standard normal distribution and by rotational symmetry of the standard normal distribution, it follows:

$$\mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] = \frac{n}{p} \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2.$$

Putting equation above in equation 10 into, we obtain

$$\mathbb{E}_{\mathbf{X}} [\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] = \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 (1 - \frac{n}{p}) + \mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2].$$

We are going to evaluate $\mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2]$. First, we observe that

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{z} - \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}} \\ &= \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}} + \mathbf{X}_{\mathcal{T}^c} \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + \boldsymbol{\epsilon} - \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}} \\ &= \mathbf{X}_{\mathcal{T}^c} \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + \boldsymbol{\epsilon}, \end{aligned} \quad (11)$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_n]^T$. Because $\mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}}$ and $\mathbf{X}_{\mathcal{T}^c} \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + \boldsymbol{\epsilon}$ are both uncorrelated, we have

$$\mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2] = \text{Tr}(\mathbb{E}_{\mathbf{X}} [\mathbf{X}_{\mathcal{T}}^{\perp T} \mathbf{X}_{\mathcal{T}}^{\perp}] \mathbb{E}_{\mathbf{X}} [\boldsymbol{\eta} \boldsymbol{\eta}^T]).$$

As $p > n$, we have $\mathbf{X}_{\mathcal{T}}^{\perp} = \mathbf{X}_{\mathcal{T}}^T (\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T)^{-1}$ and thus

$$\mathbb{E}_{\mathbf{X}} [\mathbf{X}_{\mathcal{T}}^{\perp T} \mathbf{X}_{\mathcal{T}}^{\perp}] = \mathbb{E}_{\mathbf{X}} [(\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T)^{-1}]$$

$\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T$ follows a Wishart distribution: $\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T \sim \mathcal{W}_n(p, \mathbf{I}_n)$, and $(\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T)^{-1}$ follows an inverse Wishart distribution: $(\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T)^{-1} \sim \mathcal{W}_n^{-1}(p, \mathbf{I}_n)$. Its expectation is given by:

$$\mathbb{E}[(\mathbf{X}_{\mathcal{T}} \mathbf{X}_{\mathcal{T}}^T)^{-1}] = \frac{\mathbf{I}_n}{p-n-1}.$$

We have

$$\mathbb{E}_{\mathbf{X}} [\|\mathbf{X}_{\mathcal{T}}^{\perp} \boldsymbol{\eta}\|_2^2] = \frac{1}{p-n-1} \mathbb{E}_{\mathbf{X}} [\boldsymbol{\eta}^T \boldsymbol{\eta}].$$

From equation 11, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{X}} [\boldsymbol{\eta}^T \boldsymbol{\eta}] &= \mathbb{E}_{\mathbf{X}} [(z - \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}})^T (z - \mathbf{X}_{\mathcal{T}} \mathbf{w}_{\mathcal{T}}^{\text{OOD}})] \\ &= \mathbb{E}_{\mathbf{X}} [(\mathbf{X}_{\mathcal{T}^c} \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + \boldsymbol{\epsilon})^T (\mathbf{X}_{\mathcal{T}^c} \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + \boldsymbol{\epsilon})] \\ &= [\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}]^T \mathbb{E}_{\mathbf{X}} [\mathbf{X}_{\mathcal{T}^c}^T \mathbf{X}_{\mathcal{T}^c}] \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + \underbrace{\mathbb{E}_{\mathbf{X}} [\boldsymbol{\epsilon}^T]}_{=0} \underbrace{\mathbb{E}_{\mathbf{X}} [\mathbf{X}_{\mathcal{T}^c}]}_{=0} \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + [\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}]^T \underbrace{\mathbb{E}_{\mathbf{X}} [\mathbf{X}_{\mathcal{T}^c}^T]}_{=0} \underbrace{\mathbb{E}_{\mathbf{X}} [\boldsymbol{\epsilon}]}_{=0} + \mathbb{E}_{\mathbf{X}} [\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}]. \end{aligned}$$

$\mathbf{X}_{\mathcal{T}^c}^T \mathbf{X}_{\mathcal{T}^c}$ follows a Wishart distribution, *i.e.*, $\mathbf{X}_{\mathcal{T}^c}^T \mathbf{X}_{\mathcal{T}^c} \sim \mathcal{W}_q(n, \mathbf{I}_q)$. We obtain thus

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}[\boldsymbol{\eta}^T \boldsymbol{\eta}] &= n[\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}]^T \mathbf{w}_{\mathcal{T}^c}^{\text{OOD}} + n\sigma^2 \\ &= n\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + n\sigma^2.\end{aligned}\tag{12}$$

From equation 12, we deduce that

$$\mathbb{E}_{\mathbf{X}}[\|\mathbf{X}_{\mathcal{T}}^+ \boldsymbol{\eta}\|_2^2] = \frac{n}{p-n-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2).$$

and

$$\mathbb{E}_{\mathbf{X}}[\|\hat{\mathbf{w}}_{\mathcal{T}} - \mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2] = \|\mathbf{w}_{\mathcal{T}}^{\text{OOD}}\|_2^2 (1 - \frac{n}{p}) + \frac{n}{p-n-1} (\|\mathbf{w}_{\mathcal{T}^c}^{\text{OOD}}\|_2^2 + \sigma^2).$$

□

B Details about Baselines and Architecture Implementations

B.1 Baseline OOD Methods

In this section, we present an overview of the baseline methods used in our experiments. We describe the principles behind these baselines, and the chosen hyperparameters. It is worth noting that extensive hyperparameter search for each method were not performed to maintain stability. Hence, once the final model is selected, hybrid methods like ViM, ASH and NECO performance may increase if such task is performed.

Softmax Score. This score uses the maximum softmax probability (MSP) of the model as an OOD scoring function (Hendrycks & Gimpel, 2017).

Energy. Liu et al. (2020) proposes using the energy score for OOD detection, where the energy function maps the logit outputs to a scalar. To maintain the convention that lower scores correspond to in-distribution (ID) data, (Liu et al., 2020) uses the negative energy as the OOD score.

ReAct. Sun et al. (2021) propose clipping extreme-valued activations. The original paper found that clipping activations at the 90th percentile of ID data was optimal. Moreover, as the authors propose, we employ the ReAct+Energy configuration.

KL-Matching & MaxLogit. KL-Matching computes the class-wise average probability using the entire training dataset. Consistent with the approach outlined in (Hendrycks et al., 2022), this calculation is based on the predicted class rather than the ground-truth labels. MaxLogit employs the maximum logit value of the model as an OOD scoring function.

Mahalanobis. This score leverages the feature vector from the layer preceding the final classification layer (Lee et al., 2018a). To estimate the precision matrix and the class-wise mean vector, we used the entire training dataset. It's important to note that we incorporated ground-truth labels during this computation process.

ViM & Residual. Wang et al. (2022) decomposes the latent space into a principal space P and a null space $P^\perp$. The ViM score is calculated by projecting the features onto the null space to create a virtual logit, which is then combined with the logits using the norm of this projection. To enhance performance, they calibrate this norm with a constant which is determined by dividing the sum of the maximum logits by the sum of the norms of the null space projections, both measured on the training set. The Residual score is derived by computing the norm of the latent vector's projection onto the null space. We followed the author's suggestions for the null space, by setting it to half the size of the full feature vector, adapted to each model width.

ASH. Djuricic et al. (2023) employs activation pruning at the penultimate layer, just before the application of the DNN classifier. This pruning threshold is determined on a per-sample basis, eliminating the need for pre-computation of ID data statistics. The original paper presents three different post-hoc scoring functions, with the only distinction among them being the imputation

method applied after pruning. We employ ASH-P in our experiments as it performed the best, in which the clipped values are replaced with zeros. As specified in the original paper, we fix the pruning threshold value to 90%.

NECO. Ammar et al. (2024) leverages the geometric properties of Neural Collapse, measuring the relative norm of a sample within the subspace defined by the ETF to identify OOD samples. NC typically involves a collapse in the variability of class representations, leading to a more structured and simplified feature space. It is hypothesized that this collapse also impacts OOD detection, particularly through the emerging orthogonality between ID and OOD data. NECO utilizes this orthogonality to effectively distinguish between ID and OOD data by measuring the relative norm of each data point within the approximated ETF space scaled by the maximum logit value as the OOD score. We use a dimension $d = c$ to approximate the ETF sub-space for all architectures, with c being the number of classes.

B.2 Experiments Setup

ViT Experimental Setup. For all experiments, we trained a set of ViT models with hidden dimensions [4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 60, 80, 100, 120, 160, 200, 240, 280, 320, 360, 400, 480, 520, 600, 680, 760, 800]. This hidden dimension serve as the dimension of the output layer after the linear transformation (the class-token size). The dimension of the FeedForward layer is the width multiplied by 4. The input size is set to 32 and the patch size to 8, no dropout is used and we use 4 heads with 4 Transformer blocks. The ViT models are first randomly initialized and then trained on CIFAR-10 using stochastic gradient descent with CE loss. The weights are fine-tuned for 60 000 steps, with no warm-up steps, 1 024 batch size, 0.9 momentum, and a learning rate of 0.03.

Swin Experimental Setup. We used a standard 4 block Swin architecture, with a downscaling factor of (2,2,2,1) for each block respectively. The width ranges from 1 to 100, with a window size of 4, an input size of 32, and a filter size of 4. The model is randomly initialized and then optimized using an Adam optimizer with CE loss for a 1 000 epoch using a batch size of 1 024. The initial learning rate is 0.0001.

CNN Experimental Setup. Similar to Nakkiran et al. (2021), we define a standard family CNN models formed by 4 convolutional stages of controlled base width $[k, 2k, 4k, 8k]$, for k in the range of $[1, 128]$, with a fully connected layer as classifier. The MaxPool is set to $[2, 2, 2, 4]$ for the four blocks respectively. For all the convolution layers, the kernel size is set to 3, stride and padding to 1.

C Details about Neural Collapse

For overparametrised model trained through the terminal phase of training (TPT), Neural Collapse (NC) phenomenon emerges, particularly in the penultimate layer and in the linear classifier of DNNs (Papayan et al., 2020; Ammar et al., 2024). NC is characterized by five main properties:

1. **Variability Collapse (NC1):** the within-class variation in activations becomes negligible as each activation collapses toward its respective class mean.
2. **Convergence to Simplex ETF (NC2):** the class-mean vectors converge to having equal lengths, as well as having equal-sized angles between any pair of class means. This configuration corresponds to a Simplex Equiangular Tight Frame (ETF).
3. **Convergence to Self-Duality (NC3):** in the limit of an ideal classifier, the class means and linear classifiers of a neural network converge to each other up to rescaling, implying that the decision regions become geometrically similar and that the class means lie at the centers of their respective regions.
4. **Simplification to Nearest Class-Center (NC4):** The network classifier progressively tends to select the class with the nearest class mean for a given activation, typically based on standard Euclidean distance.
5. **ID/OOD Orthogonality (NC5):** As the training procedure advances, OOD and ID data tend to become increasingly more orthogonal to each other. In other words, the clusters of

OOD data become more perpendicular to the configuration adopted by ID data (*i.e.*, the Simplex ETF).

These NC properties provide valuable insights into DNNs learned representation structure and properties, which allows for a considerable simplification. Additionally, the convergence of NC can be linked to OOD detection Ammar et al. (2024); Haas et al. (2023); Zhao & Cao (2023); Zhang et al. (2024). For further details refer to (Papayan et al., 2020; Ammar et al., 2024)

D Complementary Results on Double Descent and OOD Detection

D.1 ReLU Random Feature Model Experimental Setting

ReLU Random Feature (RRF) models (Rahimi & Recht, 2007) offer a class of simple yet expressive models. These models approximate two-layer neural networks by fixing the first-layer weights randomly and learning linear output weights. Formally, the model is defined as:

$$RRF(x) = \sum_{i=1}^m a_i \phi(w_i^\top x + b_i), \quad \text{where } \phi(z) = \max(0, z), \quad (13)$$

with $w_i \sim \mathcal{D}_w$, $b_i \sim \mathcal{D}_b$ drawn i.i.d. from fixed distributions, and the output weights $a_i \in \mathbb{R}$ learned via linear regression on the transformed data. At their core, these models perform a randomized nonlinear projection followed by a linear mapping, offering a computationally tractable means of studying nonlinear learning behaviors.

We stack a sigmoid activation function on the model outputs, which aligns with the theoretical setting underlying Theorem 1 and Theorem 2:

$$f(x) = \phi'(RRF(x)), \quad \text{where } \phi'(z) = \frac{1}{1 + e^{-z}},$$

This allows us to closely mirror the binary classification setting.

We vary the model width (*i.e.*, the number of random features) from 2 to 10 000 and train each model for 600 epochs. For training and evaluation, we generate two datasets of 1 000 samples each by selecting two CIFAR-10 classes (class 0 and class 1) and fitting Gaussian Mixture Models (GMMs) to this data. These GMMs then serve as generative sources for the training and test samples. Out-of-distribution (OOD) data is generated by sampling 1 000 inputs from a distinct Gaussian distribution with the same shape as CIFAR-10 images.

To evaluate OOD detection, we employ the following score, which is tailored to the binary classification context:

$$g(x) = \frac{|\hat{y} - 0.5|}{0.5} \quad (14)$$

where $\hat{y}$ denotes the model's prediction. This score reflects our OOD detection risk proxy defined in equation 5, capturing the model's confidence by penalizing under-confident predictions (a sigmoid value close to 0.5). Interestingly, the AUC drops to around 50% at the interpolation threshold, indicating behavior no better than random guessing. To better understand how the OOD score evolves across the double descent curve, Figure 3 shows the distribution of $g(x)$ from equation 14 across different model complexities—specifically in underparameterized, overparameterized, and interpolation-threshold regimes. At the interpolation threshold, the distribution is notably heavy-tailed, revealing that the model is producing extreme and unjustified confident outputs, often close to 0 or 1 for all inputs. This overconfidence, applied indiscriminately to both ID and OOD samples, explains the poor AUC for the defined OOD score in equation 14. In contrast, models outside the interpolation region exhibit more informative score distributions: OOD samples tend to receive low scores, while ID scores are more uniformly spread, enabling better separation. However, as model complexity approaches the interpolation threshold from either side (width of value 500 and 5 000), score separability start diminishing, primarily because OOD samples begin receiving increasingly higher confidence scores.

D.2 CIFAR-10 Additional Results

To further show the consistency of double descent for OOD detection, Figures 4, 5 6, 7 and 8 show the OOD detection metrics performance on six more semantic-shift OOD datasets. To illustrate the

performance of other OOD-methods while maintaining visibility, we show two different methods at each dataset alongside the better-performing and most stable three: MSP, NECO, and ASH.

D.3 CIFAR-100 Results

In this section, we present results for the CIFAR-100 dataset as ID for a ResNet-18 model. Figure 9 illustrates the OOD detection metrics performance, Figure 10 shows the accuracy and eigenvalues distribution (see section E.1 for discussion about eigenvalues). We can observe similar behaviors between the ResNet-18 trained on CIFAR-10, and this current configuration on a harder dataset (CIFAR-100).

Table 2 illustrates the evolution of AUC between the underparametrized and overparametrized regime and its correlation with the $NC1_{u/o}$ for the remaining OOD datasets. As the CNN’s $NC1_{u/o} < 1$, its performance stagnates or deteriorates with overparametrization, while the other models improve. Additionally, we can see how the hybrid-based methods improve considerably and become competitive with logit-methods when $NC1_{u/o} > 1$.

Table 2: Models performance in terms of AUC in the underparametrized local minima (AUC_u) and the overparametrized maximum width (AUC_o), *w.r.t* $NC1_{u/o}$ value. Best is highlighted in **green** when AUC_u is higher, **red** when AUC_u is higher and **blue** if both AUC are within standard deviation range. The highest AUC value per-dataset and per-architecture is highlighted in **bold**.

Model	$NC1_{u/o}$	Method	ImageNet-O		Textures		iNaturalist	
			$AUC_u \uparrow$	$AUC_o \uparrow$	$AUC_u \uparrow$	$AUC_o \uparrow$	$AUC_u \uparrow$	$AUC_o \uparrow$
CNN	0.88	Softmax score	71.60±0.37	71.78±0.57	70.76±0.99	74.80±0.78	66.19±1.27	72.07±1.56
		MaxLogit	68.91±0.96	65.94±0.51	62.01±3.75	45.56±1.06	63.38±1.89	61.58±2.07
		Energy	65.06±1.57	65.76 ±0.51	53.62±5.87	45.07±1.04	60.11±2.32	61.32±2.05
		Energy+ReAct	59.06±2.60	58.08±0.80	42.80±7.54	32.67±0.86	51.17±2.82	49.41±2.47
		NECO	67.45±1.14	68.89±0.35	56.54±3.83	55.75±3.63	59.79±1.96	68.52±3.60
		ViM	67.13±1.61	65.76±0.51	49.84±3.59	45.09±0.79	51.58±2.82	61.37±1.96
		ASH-P	66.20±1.48	66.26±0.57	56.07±5.94	46.23±1.07	60.73±2.83	61.52±2.11
ResNet	1.96	Softmax score	70.90±0.52	75.91±0.49	68.03±0.75	72.78 ±1.14	65.79±2.16	74.85±1.39
		MaxLogit	69.45±0.88	72.50±0.88	65.06±1.80	64.43±2.04	63.88±3.47	72.82±2.32
		Energy	67.36±1.26	72.44±0.89	62.28±2.55	64.35±1.84	61.73±4.97	72.77±2.33
		Energy+ReAct	68.02±1.41	72.00±0.81	64.15±2.68	66.13±1.20	61.68±6.83	71.85±2.60
		NECO	70.42±0.83	76.11±1.42	67.56±1.84	73.18±3.20	64.50±3.51	75.12±2.21
		ViM	70.94±1.54	75.03±0.62	77.88±1.66	81.02±1.42	62.56±4.91	67.12±2.44
		ASH-P	67.36±1.26	71.57±0.94	62.28±2.55	62.96±2.11	61.73±4.97	71.72±2.28
Swin	1.70	Softmax score	56.62±3.15	64.78±1.48	49.71±3.38	63.51±0.33	49.10±6.81	60.26±2.08
		MaxLogit	55.86±3.95	64.70±2.07	49.47±4.18	60.29±0.38	49.22±5.07	58.91±1.81
		Energy	49.58±5.48	64.56±2.05	49.49±6.73	59.96±0.46	51.61±7.96	58.73±1.75
		Energy+ReAct	49.81±5.51	65.43±2.09	50.66±6.07	62.38±0.24	51.87±7.51	59.39±1.83
		NECO	57.63±4.14	68.19±2.71	56.50±3.64	68.06±0.66	49.09±5.14	62.58±2.21
		ViM	65.18±1.83	73.45±2.25	84.47±1.46	78.67±1.65	67.86±2.44	63.83±2.63
		ASH-P	49.46±5.66	64.48±2.11	48.96±7.85	59.90±0.48	51.33±10.79	58.69±1.99
ViT	2.32	Softmax score	64.17±0.64	63.64±0.80	67.73±1.82	70.27±0.36	52.79±0.77	58.23±0.51
		MaxLogit	63.15±0.88	68.87±0.52	67.22±2.25	79.25±0.38	51.90±1.95	61.28±0.94
		Energy	61.24±1.08	69.10±0.50	65.59±2.78	79.68±0.38	51.11±3.04	61.40±0.97
		Energy+ReAct	61.30±1.21	69.09±0.49	65.64±2.86	79.68±0.38	51.63±4.61	61.39±0.98
		NECO	65.83±1.35	69.95±0.37	69.41±2.16	77.32±0.49	52.50±2.05	62.96±0.89
		ViM	68.76±1.52	67.82±0.55	65.58±2.69	74.13±0.38	56.41±3.81	60.42±0.90
		ASH-P	61.24±1.08	68.96±0.48	65.59±2.78	79.39±0.42	51.11±3.04	61.28±0.99

D.4 ResNet-34 Results

In this section, we present the results for the ResNet-34 architecture, a deeper version of the ResNet family. Results include both CIFAR-10 and CIFAR-100 datasets as ID. Figure 12 shows the accuracy curve and Figure 11 depicts the OOD detection metrics performance. We observe a similar curve to that of ResNet-18 for both datasets and for all OOD methods, with slightly higher performances. We highlight that the interpolation threshold occurs at a smaller width for ResNet-34 ($k=8$), compared to

ResNet-18 ($k=10$) for the CIFAR-10 case. This higher complexity also contributes to its lower NC1 values in Figure 22.

D.5 Out-of-Distribution Risk Experiments

In Section 4.1, we introduced $R_{\text{OOD}}(\hat{f})$ as a metric for OOD detection. Figures 13, 14, 15, 16, and 17 show the expected OOD risk as a function of the model width and highlight a similar double descent phenomenon across all OOD datasets.

D.6 Experiments Without Label-Noise Results

Like experiments depicted in Nakkiran et al. (2021), our experimental results in Section 5 consider noisy labels. In this section, we consider a noiseless setting with Figure 18 that depicts the model accuracy as a function of the model width in a noiseless setup. Instead of a double descent phenomenon, a plateau or stagnation appears near the interpolation threshold. Similarly, Figure 19 features the OOD detection performance with respect to the model width under this setup. The curves exhibit a similar behavior than for the model accuracy. Moreover, we observe that removing label noise makes the learning task easier with better performance, e.g., ResNet-18 max accuracy rising from 83.40% to 94.48% with a similar rise observed for AUC.

E Model representations and Neural Collapse Analysis

E.1 Structure of the Model Representation

Convergence towards Neural Collapse can be an indicator of improved model representation, as defined by the ETF structure. As such, the model eigenvalues distribution can be seen as describing how much the model representation aligns with this structured manifold. The properties of the ETF implies that the top c eigenvalues will be equally prominent, as they represent the top c eigenvectors that constructs the ETF subspace (Papayan et al., 2020; Ammar et al., 2024). The remaining eigenvalues should be less influential, as the ID the data lies essentially in the c -dimensional ETF subspace. Figure 20 shows the distribution of each model eigenvalues at the overparametrized regime. It is worth noting that it seems that for all architectures, all overparametrized model widths show similar curves. From Figure 20 we can see that ResNet-18 and Swin models follow the expected NC pattern, with a steep drop in importance at the c^{th} eigenvalue indicating the limit of the ETF. However, Figure 20 also shows that ViT and CNN exhibit a slowly decaying curve, which indicates a lack of clear separation in the model representation between highly important ID information and noisy features. This lack of a global structure in their representation results in both models failing to reliably outperform their underparametrized minima.

This highlights the importance of the ETF structure and NC convergence in enhancing representation stability, which can be useful for improving ID classification and OOD detection tasks.

E.2 Evolution of NC1

In this section, we further analyze the evolution of NC1 and the model's learned representation with overparametrization, to further show their correlation. Our analysis will focus primarily on the ResNet and CNN models, due to their similarities. To compare the NC convergence across architectures, we include the overparametrized and underparametrized NC1 value of each architecture in Table 3. We will not address the transformer-based models whose performance, especially for generalization, were lower than those of ResNet or CNN. This is because transformers typically require extensive pre-training, particularly for small datasets, and this was not the case for our experiments.

In order to visualize the variability collapse predicted by NC1, Figure 21 shows the last-layer activations for both models at their optimal underparameterized and overparameterized widths. In ResNet, transitioning to overparameterized models leads to significant improvements in the compactness and separation of ID clusters, as well as enhanced orthogonality with OOD points. In contrast, the CNN model does not show clear improvements in ID compactness or OOD separation, making it difficult to determine which representation is better.

Figure 22 shows NC1 metric against the model widths for both ResNets and CNN. While both models exhibit a double descent pattern, CNN barely matches its underparameterized metric value, whereas ResNet continuously improves with added complexity. This discrepancy in NC convergence with overparametrization explains why only ResNet benefits from increased complexity, suggesting that without improvement in NC, increasing model complexity provides no benefits for the learned representations.

Table 3: Comparison of NC1 values for different architectures in under- and over-parameterized regimes.

Architecture	NC1 (Under-parameterized)	NC1 (Over-parameterized)
ResNet-18	1.35 ± 0.04	0.687 ± 0.03
ResNet-34	1.29 ± 0.02	0.64 ± 0.02
CNN	2.37 ± 0.08	2.71 ± 0.03
Swin	35.08 ± 3.53	20.65 ± 1.88
ViT	29.52 ± 2.17	12.71 ± 0.74
ResNet (Depth)	1.53 ± 0.02	1.57 ± 0.09

E.3 Mahalanobis and Residual Joint Performance

We noticed that for all architectures, and on all OOD datasets, Mahalanobis and Residual follow the same evolution curve, with usually slightly higher *AUC* in favor of Mahalanobis. This behavior is intriguing, due to the fact that each method relies on different types of information. While Mahalanobis models the ID distribution, *i.e.*, the principal space, Residual relies on computing the null space norm, which is orthogonal to the principal space.

We associate this behavior with the noise isolation in each architecture, which is specific to the double descent training paradigm. Indeed, in order for models to be able to perfectly interpolate all the training data and achieve (almost) zero training error, noisy samples must be represented closer to their assigned (noisy) label, rather than to their true label. This will cause the train class clusters (using the true labels) to be less compact and separable, making their high-likelihood region to span almost the entire principal space, in which the ID data is represented. Hence, to separate ID from OOD, learning the Mahalanobis GMM (fitted on the train data) becomes equivalent to separating the principal and null space, which is the same reasoning behind the Residual score.

This overfitting occurs at the interpolation threshold, which causes the learned distributions by Mahalanobis to be sparse and not robust to OOD data, impeding its improvement as we transition towards overparametrization. It is important to note also that both of these methods are usually below, or struggle to surpass the random choice threshold of 0.5 *AUC* in the overparametrized regime (with the exception of texture dataset on ResNet-18 case).

Interestingly, both of these methods suffer much less from this behavior under the Transformer based architecture, and even exhibit a double descent curve on most datasets. This can be explained by the fact that even the most overparametrized Transformer variant have a training error higher than 4%. In comparison, convolutional models consistently achieve a training error lower than 0.01%. Hence, Transformers suffer less from this effect because they have not interpolated the noise in the training data perfectly. It is worth noting that interpolating the noise is desirable, as it is necessary for Generalization in this setup (Bartlett et al., 2020). Transformer-based architectures require extensive pre-training to generalize well, especially for small scale dataset, which was not performed in our experiments. This inability of transformers to perfectly interpolate the training data contributes to their lower performance in terms of Generalization in the overparametrized regime, especially in the ViT case.

F Depth-Wise Double Descent

In this section, we investigate how the model depth affects ID generalization and post-hoc OOD detection performance. Most works on the double descent phenomenon focus on varying the network width rather than depth (Belkin et al., 2020; Couillet & Liao, 2022; Bach, 2024; Nakkiran et al., 2021). A primary reason stems from challenges in modifying network depth within both empirical and theoretical frameworks. From an empirical perspective, increasing depth often introduces

training instabilities and requires hyperparameter adjustments (e.g., learning rates) that complicate comparative analyses. From a theoretical perspective, studies typically rely on frameworks like Random Matrix Theory (RMT) or statistical physics, which often model networks as linear systems with random features. Increasing the depth also increases nonlinear interactions and hierarchical dependencies, which may limit the applicability of these mathematical tools.

F.1 Experimental Setup

We conduct experiments on ResNet-like architectures composed of four residual blocks, with CIFAR-10 as the ID dataset. The width of the network is fixed (at $k = 10$), while the model complexity is controlled by modifying the model depth. In particular, the model depth is defined as the total number of residual layers across all four blocks. Starting with shallow configurations, we incrementally increase the number of layers per block, focusing on deepening of the blocks with the fewest layers first. For instance, the sequence of architectures progresses as follows:

$$1, 1, 1, 1 \rightarrow 2, 1, 1, 1 \rightarrow 2, 2, 1, 1 \rightarrow 2, 2, 2, 2 \rightarrow 3, 2, 2, 2 \rightarrow \dots \rightarrow 10, 10, 10, 10,$$

where each tuple denotes the number of residual layers in each block, respectively. In Figure 23, the x -axis depicts the total depth, i.e. the sum of the four block values. To obtain architectures with depth less than 4, we replace residual blocks with MaxPooling layers to maintain a constant overall width across all models. The training protocol follows the same setup as the width-wise experiments (see Section 5.2). Furthermore, the same set of metrics is considered: test accuracy, OOD detection AUC, Neural Collapse, and the OOD risk defined in equation 5.

F.2 Results

Figure 23 depicts a similar double descent phenomenon for model depth as observed for model width. Furthermore, Table 4 shows a correlation between the $NC1_{u/o}$ metric and OOD detection performance, which may suggest an alignment between the $NC1_{u/o}$ and OOD performance.

Table 4: Models performance in terms of AUC in the underparametrized local minima (AUC_u) and the overparametrized maximum width (AUC_o), *w.r.t* $NC1_{u/o}$ value. Best is highlighted in **green** when AUC_o is higher, **red** when AUC_u is higher and **blue** if both AUC are within standard deviation range. The highest AUC value per-dataset and per-architecture is highlighted in **bold**.

Model	$NC1_{u/o}$	Method	SUN		Places365		CIFAR-100	
			$AUC_u \uparrow$	$AUC_o \uparrow$	$AUC_u \uparrow$	$AUC_o \uparrow$	$AUC_u \uparrow$	$AUC_o \uparrow$
Depth	0.97	Softmax	72.45±0.99	69.23±2.66	72.47±0.86	69.01±0.97	71.62±0.65	68.84±1.39
		MaxLogit	72.57±0.72	71.00±2.86	72.57±1.15	70.76±1.93	70.66±0.74	70.82±2.16
		Energy	71.81±0.70	70.98±2.87	71.83±1.24	70.74±1.92	69.43±0.80	70.81±2.17
		ReAct	71.83±1.58	71.56±2.99	71.89±1.83	71.30±2.04	70.13±1.32	70.71±2.28
		NECO	72.83±0.89	71.24±2.12	72.74±1.03	70.77±1.99	70.85±0.76	69.54±1.07
		ViM	71.86±2.11	70.32±4.91	71.07±1.70	69.46±3.92	71.14±1.31	67.29±1.70
		ASH-P	71.78±0.66	70.50±3.29	71.79±1.27	70.19±2.40	69.40±0.87	70.34±2.45

F.3 Connection to Width-Wise Experiments.

Although the depth-wise experiments also exhibit a double descent phenomenon, it differs in important ways from the width-wise case observed in ResNet-18 and ResNet-34. Notably, the best models in both the underparameterized and overparameterized regimes achieve comparable performance. This distinct behavior is reflected in the ratio $NC1_{u/o} = 0.97 < 1$, indicating that feature collapse did not improve with increased depth.

These findings emphasize that out-of-distribution (OOD) performance is influenced not only by architectural biases but also by the specific training dynamics and representational geometry induced by scaling along different dimensions. It is important to note that at very large depths, additional complexities such as optimization instabilities and slower convergence emerge.

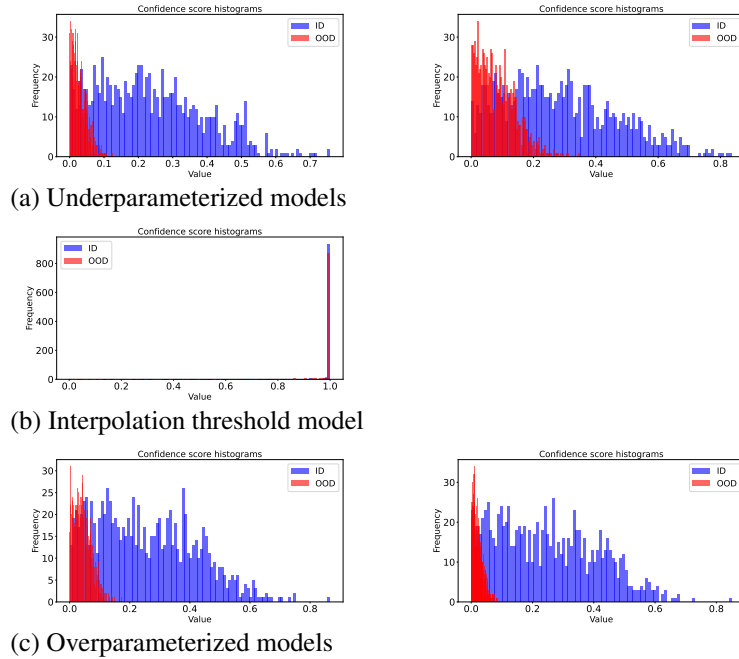


Figure 3: Distribution of the values of the OOD scoring function $g(x)$ defined in equation 14, evaluated on Random ReLU Feature (RRF) models with varying widths. (a) Score distributions for underparameterized models with widths, from left to right, of 100 and 500 respectively. (b) Score distribution at the interpolation threshold (width = 1000), where performance degrades sharply. (c) Score distributions for overparameterized models with widths, from left to right, of 5 000 and 10 000 respectively.

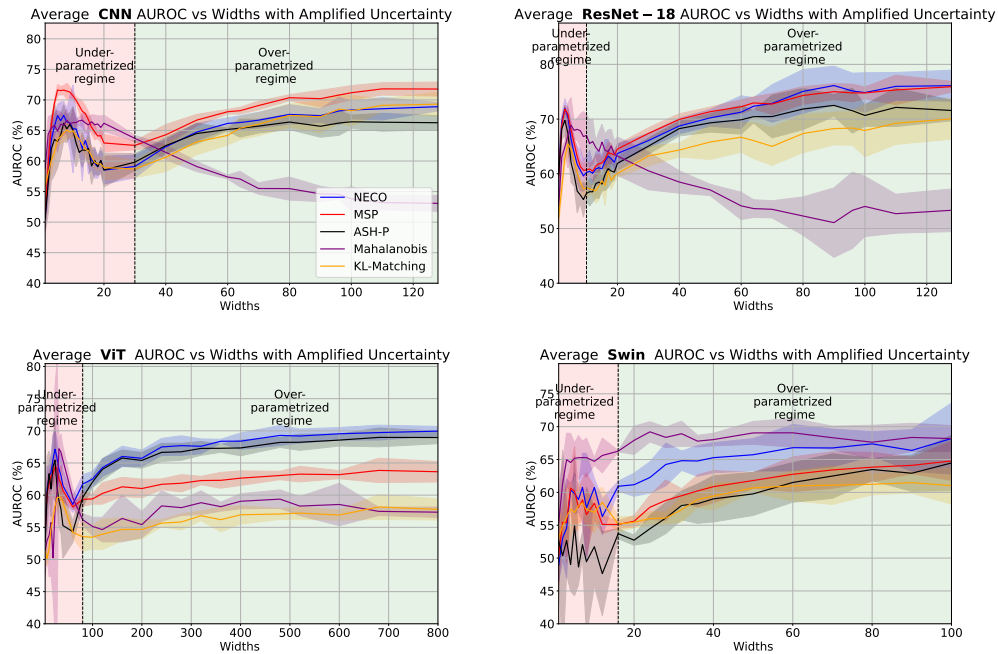


Figure 4: OOD detection (AUC) metric versus model width. Experiments performed on CNN, ResNet-18, ViT, and Swin with CIFAR10 as ID dataset and ImageNet-O as OOD dataset.

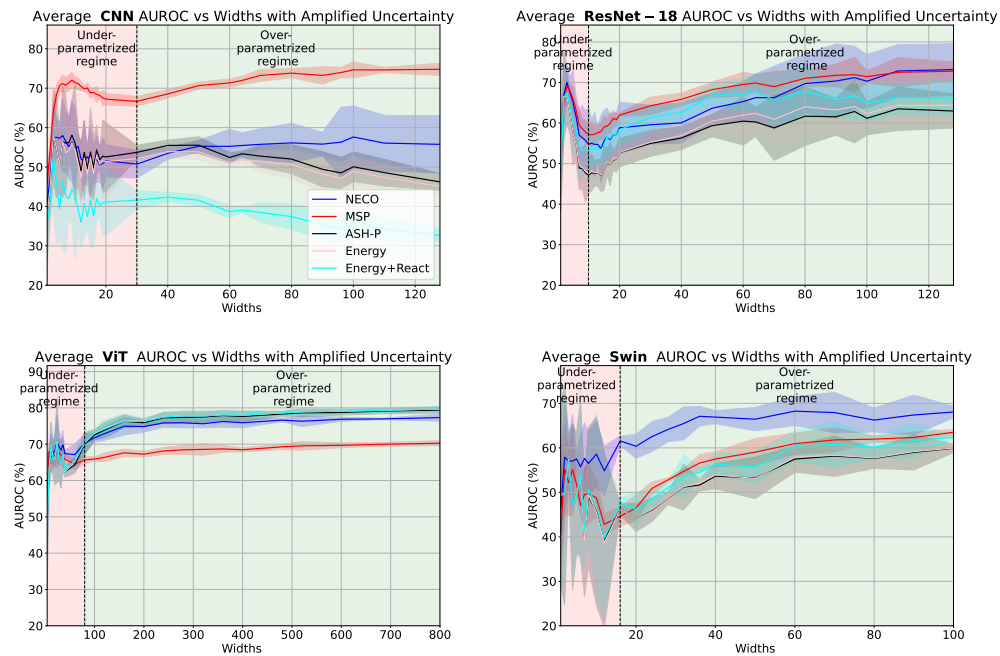


Figure 5: OOD detection (AUC) metric versus model width. Experiments performed on CNN, ResNet-18, ViT and Swin with CIFAR10 as ID dataset and Texture as OOD dataset.

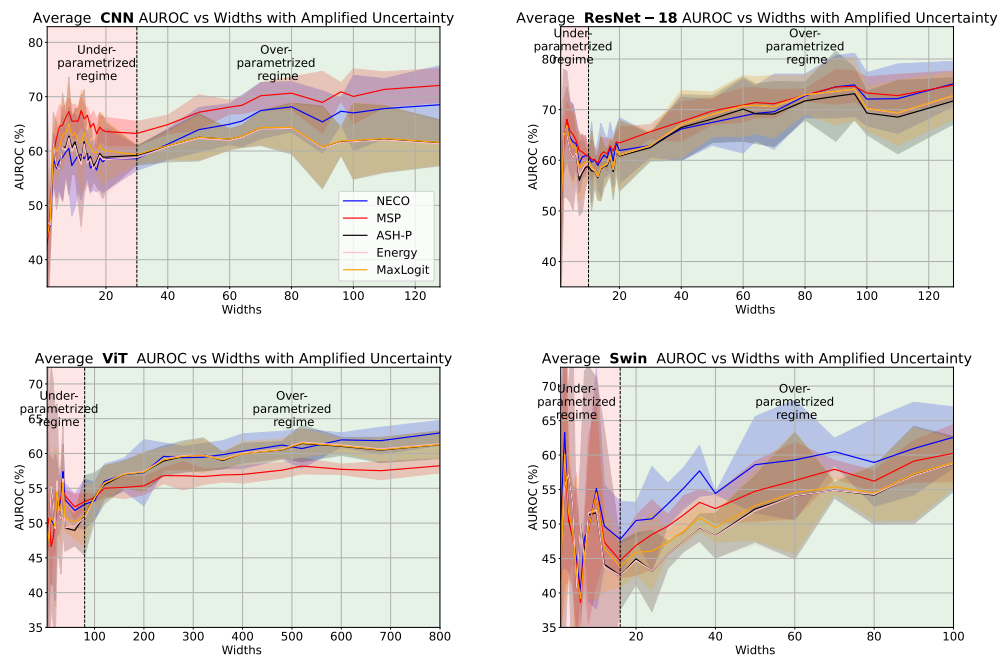


Figure 6: OOD detection (AUC) metric versus model width. Experiments performed on CNN, ResNet-18, ViT, and Swin with CIFAR10 as ID dataset and iNaturalist as OOD dataset.

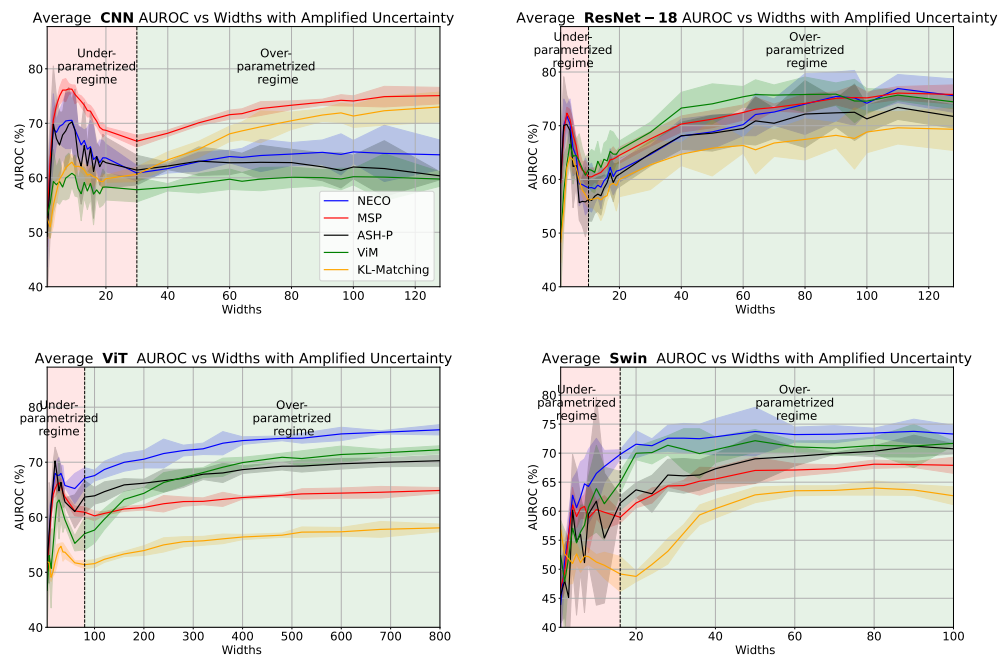


Figure 7: OOD detection (AUC) metric versus model width. Experiments performed on CNN, ResNet-18, ViT, and Swin with CIFAR10 as ID dataset and SUN as OOD dataset.

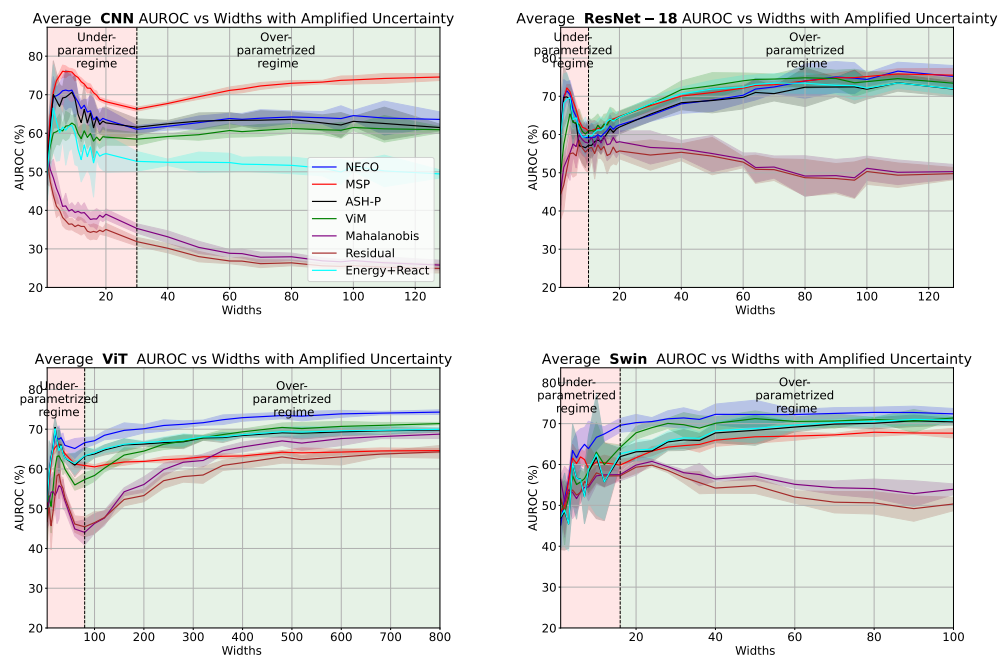


Figure 8: OOD detection (AUC) metric versus model width. Experiments performed on CNN, ResNet-18, ViT, and Swin with CIFAR10 as ID dataset and places365 as OOD dataset.

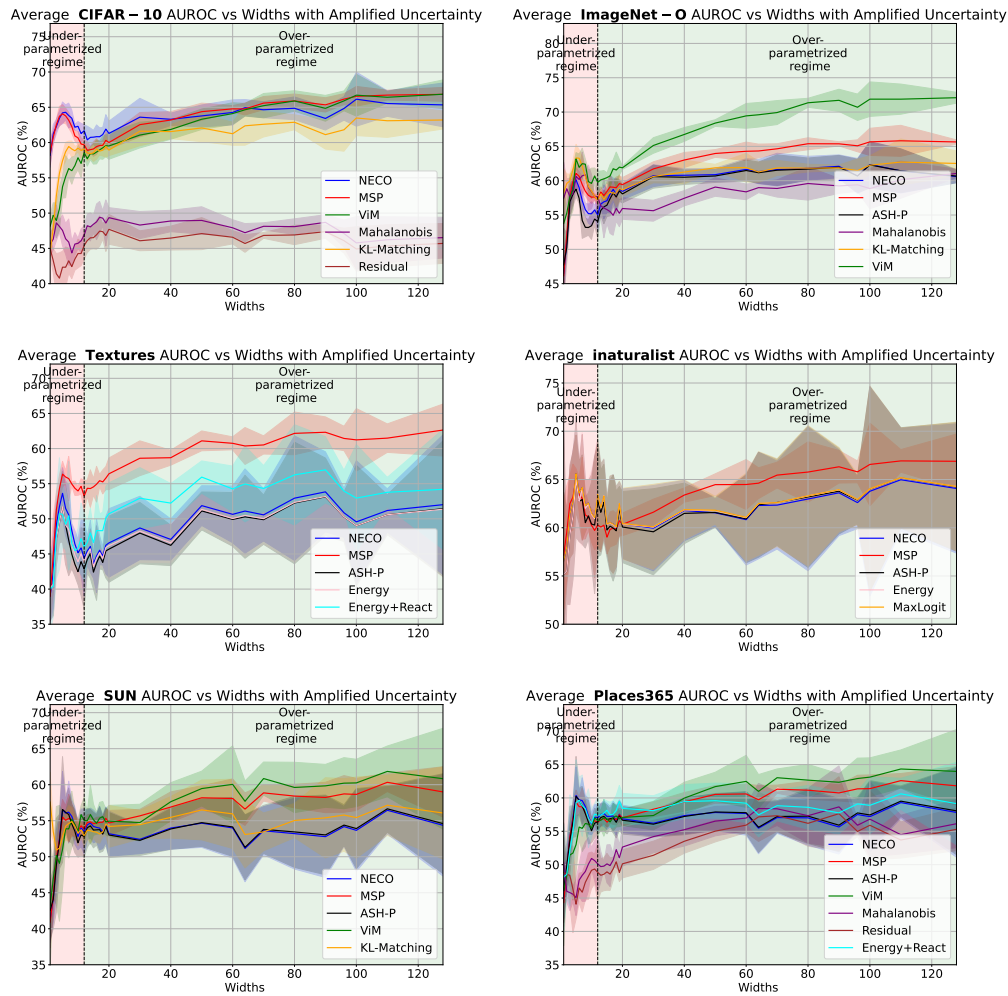


Figure 9: OOD detection (AUC) metric versus model width. Experiments performed on ResNet-18 with CIFAR100 as ID dataset and CIFAR-10, ImageNet-O, Texture, iNaturalist, SUN and places365 as OOD dataset.

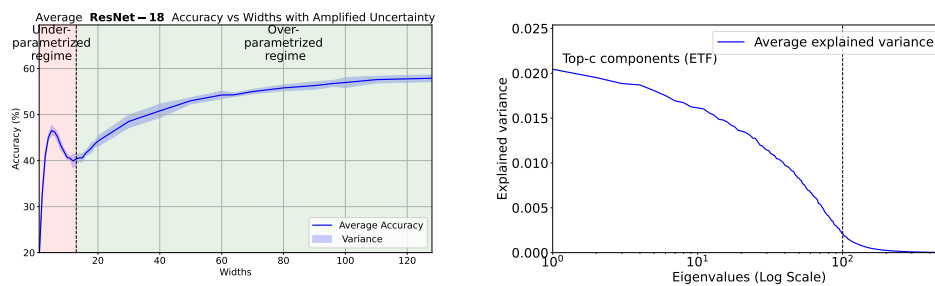


Figure 10: Accuracy versus model width (left) of ResNet-18 model on CIFAR-100. Explained variance versus eigenvalues (right) in the overparametrized regime for ResNet-18 (width 64) on CIFAR-100. The black line depicts the 100th eigenvalue.

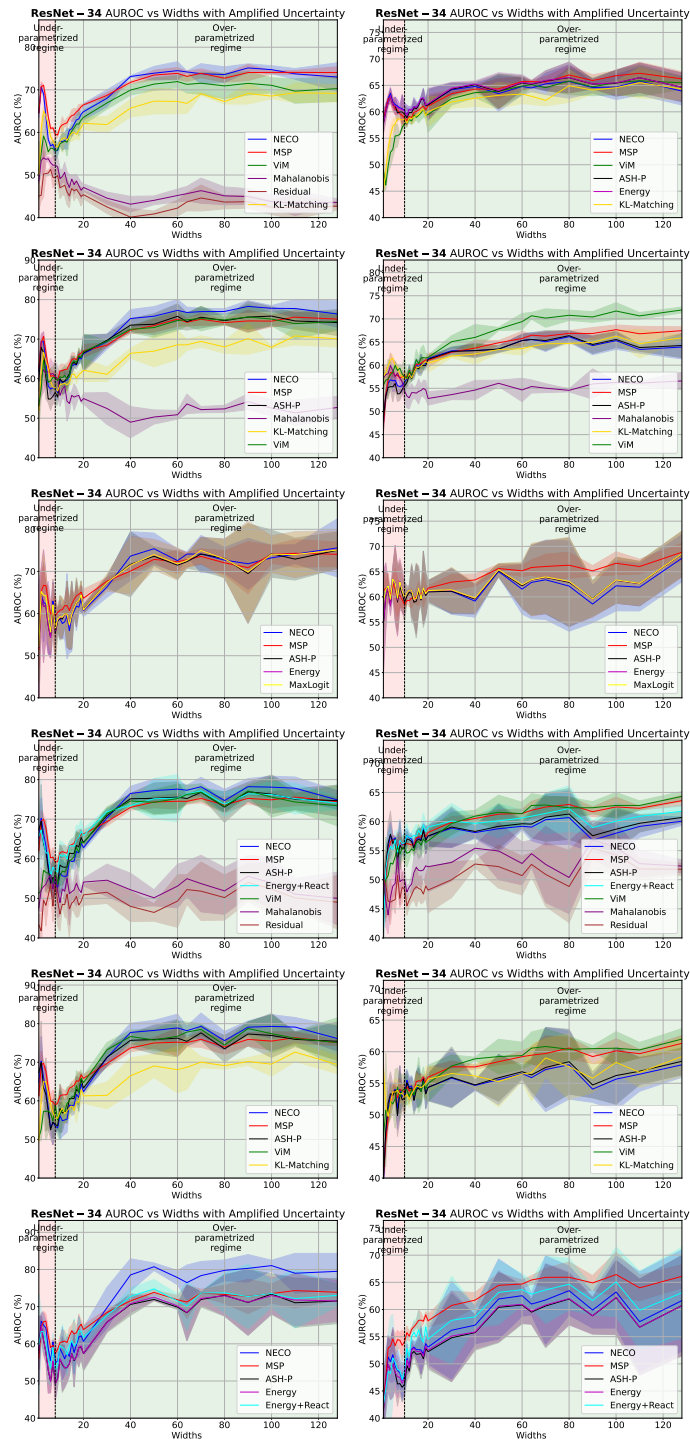


Figure 11: OOD detection (AUC) metric versus model width. Experiments performed on ResNet-34 with CIFAR10 (left) and CIFAR-100 (right) as ID dataset and CIFAR-100 (right), CIFAR-10 (left), ImageNet-O, iNaturalist, places365, SUN and Textures as OOD dataset (from top to bottom).

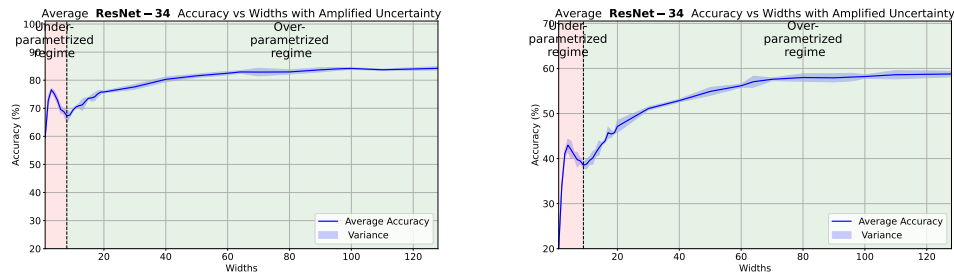


Figure 12: Accuracy versus model width. Experiments performed on CNN, ResNet-34 with CIFAR10 (left) and CIFAR100 (right) as ID datasets.

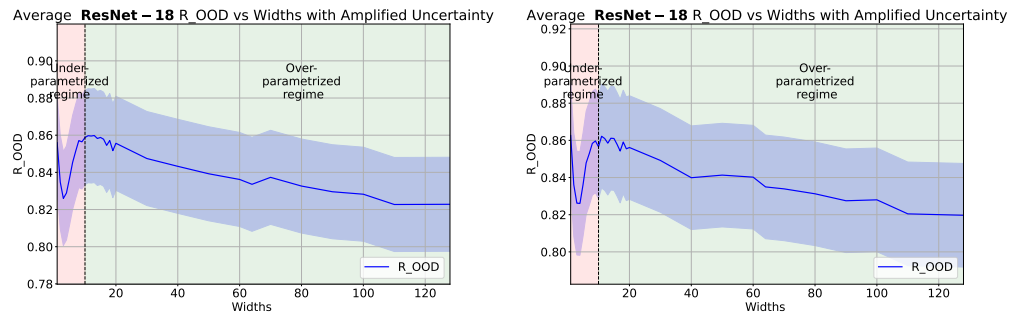


Figure 13: OOD risk metric versus model width. Experiments performed on ResNet-18 with CIFAR10 as ID dataset and CIFAR100 (left) and ImageNet-O (right) as OOD datasets.

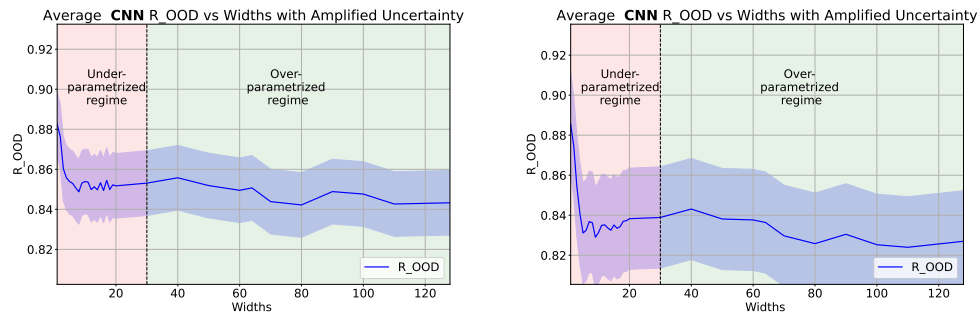


Figure 14: OOD risk metric versus model width. Experiments performed on CNN with CIFAR10 as ID dataset and iNaturalist (left) and Textures (right) as OOD datasets.

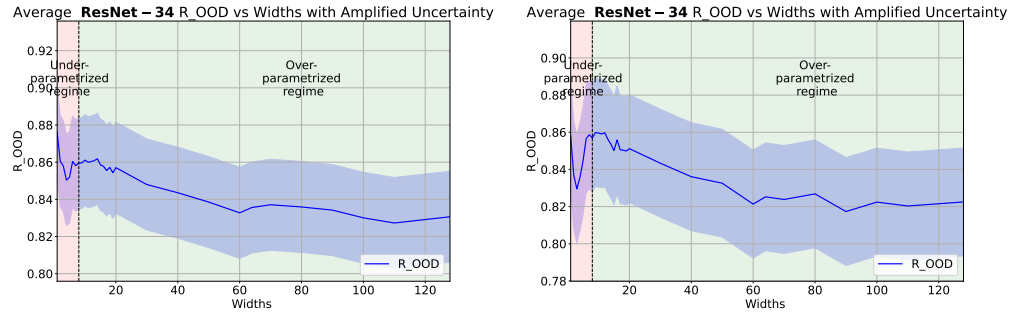


Figure 15: OOD risk metric versus model width. Experiments performed on ResNet-34 with CIFAR10 as ID dataset and iNaturalist (left) and ImageNet-O (right) as OOD datasets.

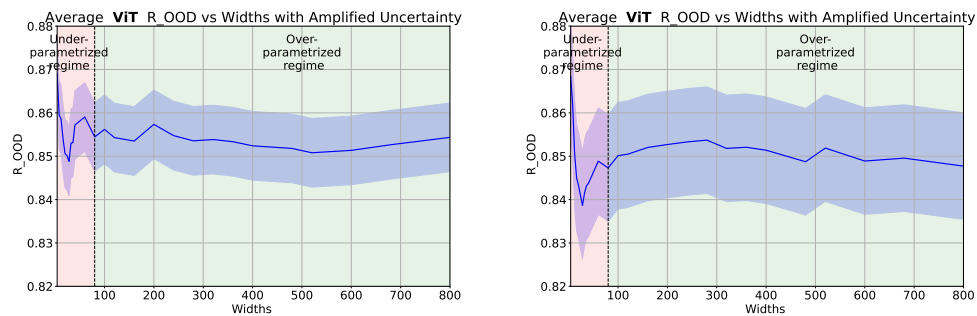


Figure 16: OOD risk metric versus model width. Experiments performed on ViT with CIFAR10 as ID dataset and Places365 (left) and ImageNet-O (right) as OOD datasets.

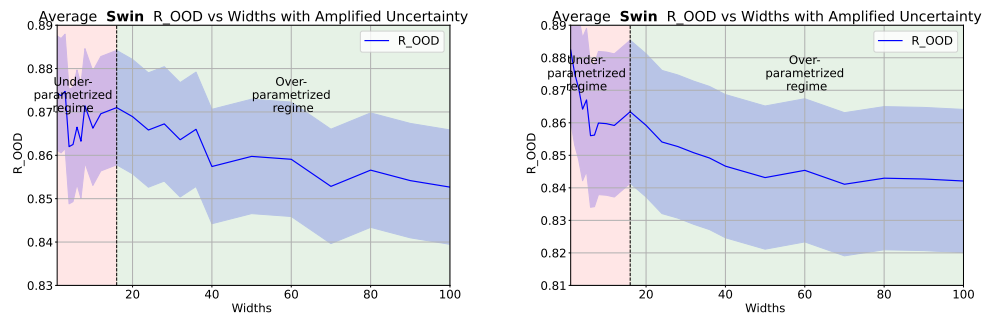


Figure 17: OOD risk metric versus model width. Experiments performed on Swin with CIFAR10 as ID dataset and SUN (left) and ImageNet-O (right) as OOD datasets.

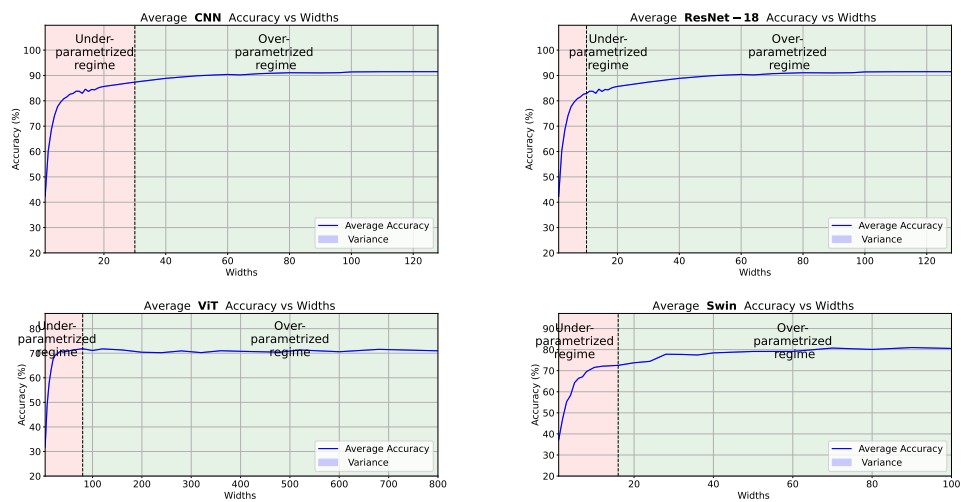


Figure 18: Accuracy versus model width. Experiments performed on CNN (top left), ResNet-18 (top right), ViT (bottom left) and Swin (bottom right) with CIFAR10 as ID dataset in the noiseless setting.

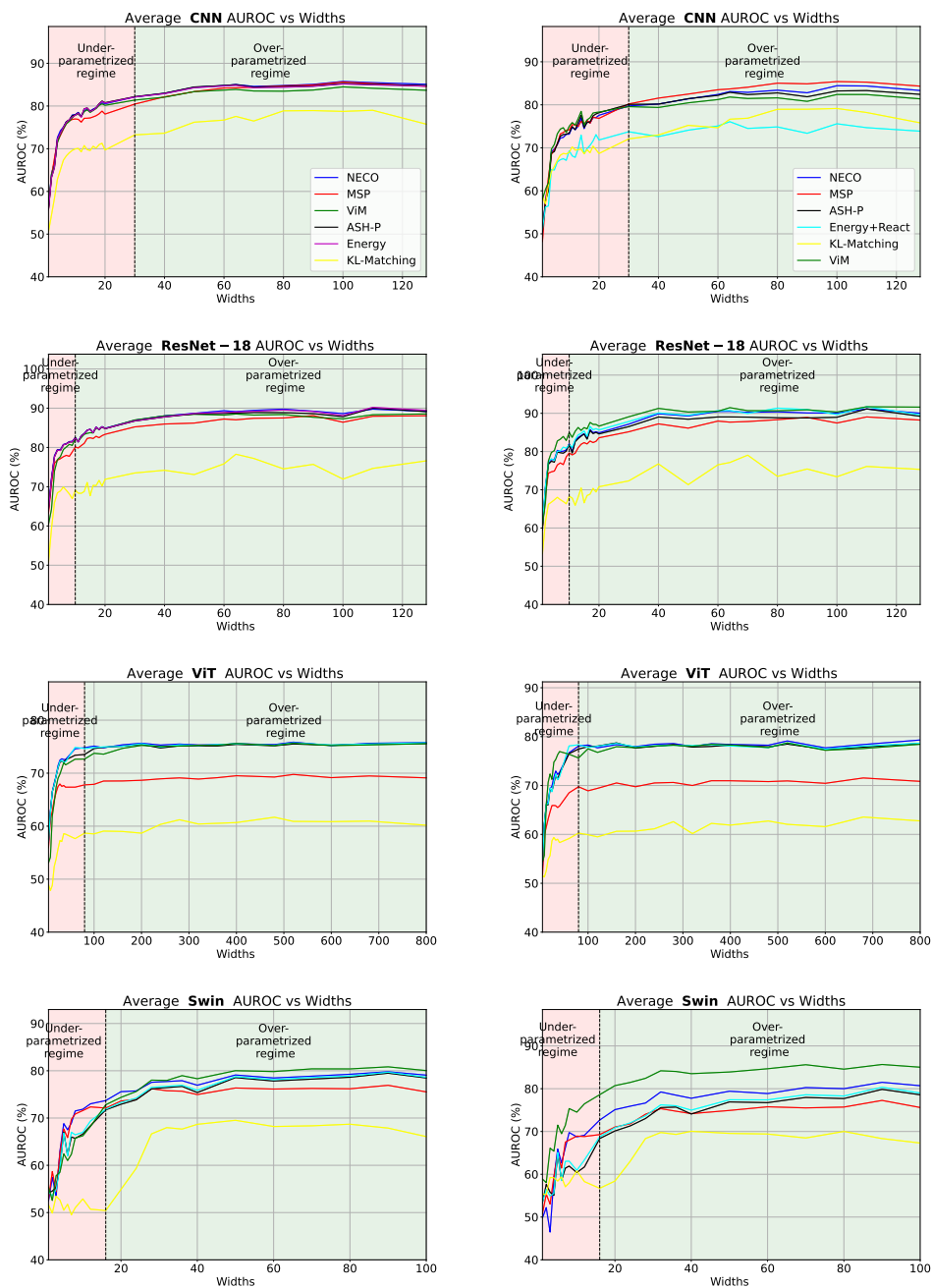


Figure 19: OOD detection (AUC) metric versus model width. Experiments performed on CNN, ResNet-18, ViT and Swin with CIFAR10 as ID dataset and CIFAR-100 (left) and ImageNet-O as (right) OOD datasets in the noiseless setting.

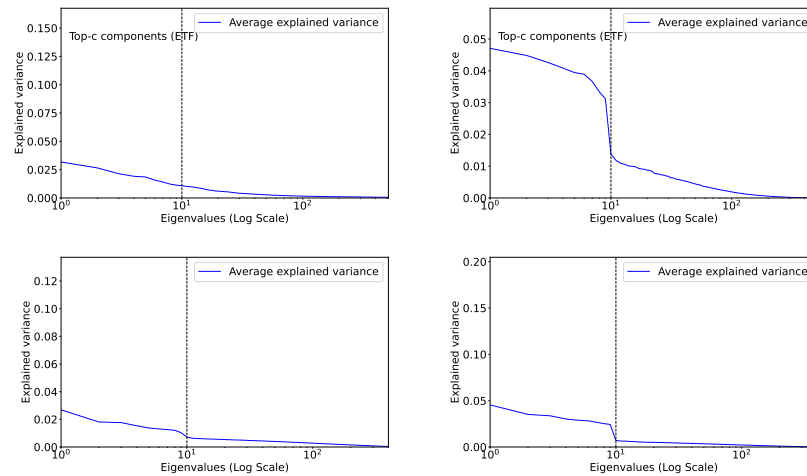


Figure 20: Explained variance versus eigenvalues in the overparametrized region for CNN (width 64), ResNet-18 (width 64), ViT (width 400) and Swin (width 50) with CIFAR-10 as ID dataset. Black line depicts the 10th eigenvalue.

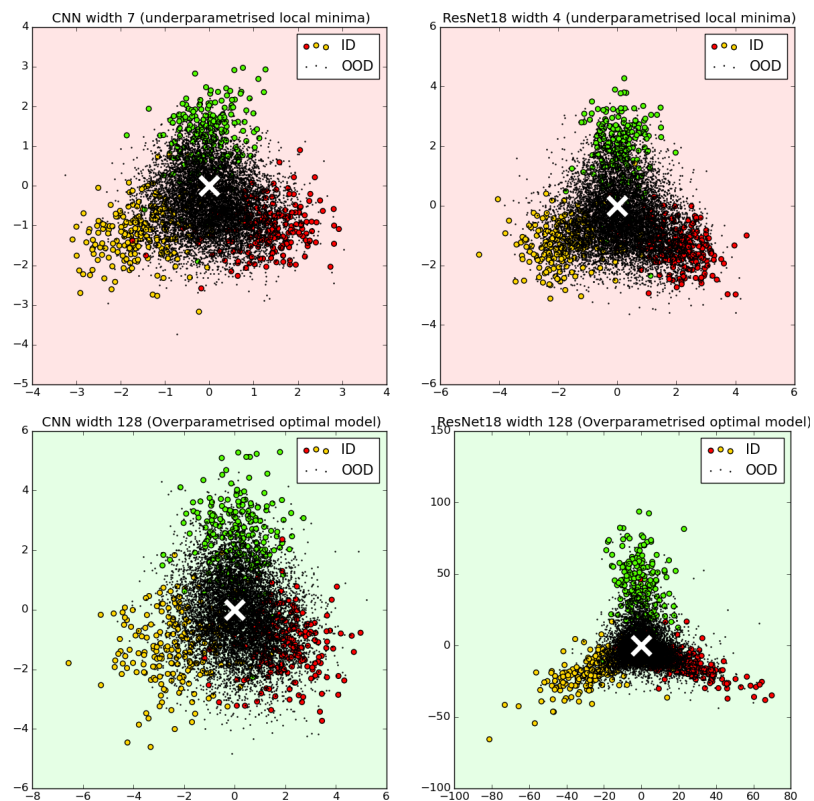


Figure 21: Visualization of the last-layer activations on the test set for ResNet and CNN in the underparametrized local minima and the overparametrized width 128 model, with cifar10 as ID and cifar100 as OOD datasets. ID points are shown in colors and OOD in black. ResNet underparametrized (top right), ResNet overparametrized (bottom right), CNN underparametrized (top left) CNN overparametrized (bottom left).

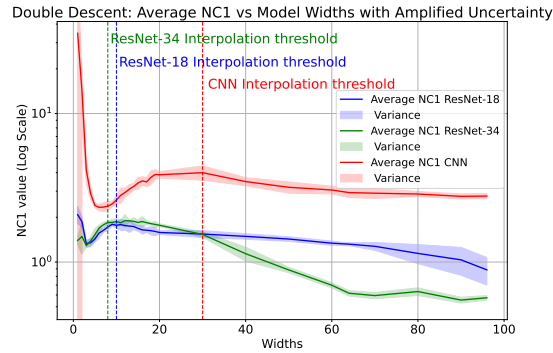


Figure 22: NC1 metrics evolution (Log scale) versus model width. ResNet-18 is shown in blue, ResNet-34 in green CNN in red. Dashed lines represent the interpolation thresholds for each model with matching color.

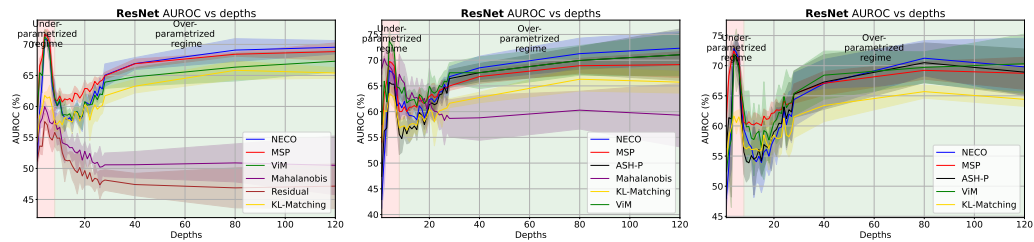


Figure 23: OOD detection (AUC) metric versus **model depth** (total number of residual layers across four blocks), with width fixed to 10. Experiments performed ResNet type models, with CIFAR10 as ID dataset and CIFAR100 (left), ImageNet-o (middle), and SUN (right) as OOD dataset.

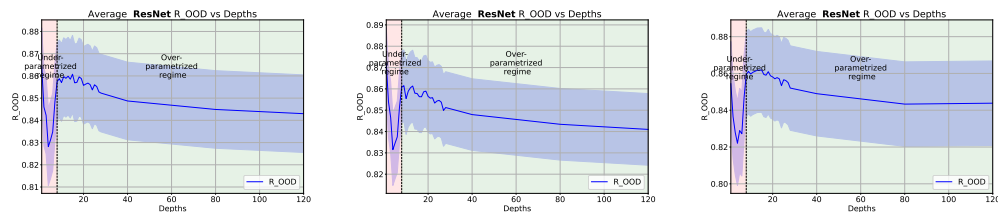


Figure 24: OOD risk metric versus **model depth**. Experiments performed on a ResNet with CIFAR10 as ID dataset and CIFAR100 (left), ImageNet-o (middle), and SUN (right) as OOD datasets.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The empirical investigation of double descent occurrence in Post-hoc OOD detection is detailed in Section 5. Theoretical insights explaining this occurrence are provided in Section 4. Section 5.7 details the suggested method to identify whether simpler models are better for OOD detection.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations of our work in our conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide the correct and complete proofs for each theoretical result, which are illustrated in detail in the Appendix A.1 and A.2. The full set of assumptions are provided in Section 4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Datasets, models, and hyperparameters used are all described in details. See Section 5 and Appendix B.2, Appendix D.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will provide the code to reproduce the experimental results in the paper upon publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All necessary hyperparameters to understand our results are detailed in Section 5 and Appendix B.2, Appendix D.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We plot the statistical variance across many seeds in our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Due to the nature of our study—exploring double descent and out-of-distribution (OOD) detection across a wide range of model architectures, dataset sizes, and OOD detection methods—we trained and evaluated a large number of models under varying conditions. Because of this variability and the exploratory scope, we did not report exact compute details (e.g., runtime, memory, or hardware) for each configuration. However, all experiments were conducted using standard GPU-equipped machines, and we aim to release code and configurations to support reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research is conducted with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets used in the paper are properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper made no such use of LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.