The Separation Capacity of Random Neural Networks

Sjoerd Dirksen
Utrecht University
s.dirksen@uu.nl

Martin Genzel
Utrecht University
m.genzel@uu.nl

Laurent Jacques
UCLouvain
laurent.jacques@uclouvain.be

Alexander Stollenwerk
UCLouvain
alexander.stollenwerk@uclouvain.be

***Abstract*—Neural networks (NNs) with random weights appear in a variety of machine learning applications, perhaps most prominently as initialization of many deep learning algorithms. We take one step closer to their theoretical foundation by addressing the following data separation problem: Under what conditions can a random NN make two classes $\boldsymbol{\mathcal{X}}^-, \boldsymbol{\mathcal{X}}^+ \subset \mathbb{R}^{\boldsymbol{d}}$ (with positive distance) linearly separable? We show that a two-layer ReLU-network with standard Gaussian weights and uniformly distributed biases can solve this problem with high probability. Crucially, the number of required neurons is explicitly linked to geometric properties of the underlying sets $\boldsymbol{\mathcal{X}}^-$ and $\boldsymbol{\mathcal{X}}^+$. This instance-specific viewpoint allows us to overcome the usual curse of dimensionality (exponential width of the layers) in non-pathological situations where the data carries low-complexity structure. The key ingredient to make this intuition precise is a geometric notion of complexity (based on the Gaussian mean width), which leads to sound and informative separation guarantees. On a technical level, our proof strategy departs from the standard statistical learning approach, and instead relies on tools from convex geometry and high-dimensional probability. Besides the mathematical results, we also discuss several connections to machine learning, such as memorization, generalization, and adversarial robustness.**

The following is a very condensed overview of our recent findings in [1]. For more details, possible refinements, related literature, and proofs, we refer to this article.

I. Introduction

We are concerned with the following fundamental problem on class separability via random NNs:

Problem 1. Consider two bounded sets $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{R}^d$ that are δ-*separated*, i.e., $\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \geq \delta$ for all $\boldsymbol{x}^+ \in \mathcal{X}^+$ and $\boldsymbol{x}^- \in \mathcal{X}^-$. Let $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$ represent a (multi-layer) feedforward NN with *random weights*, where the size of F may depend on $\mathcal{X}^-$ and $\mathcal{X}^+$.

> *Under what conditions does F make the classes $\mathcal{X}^-$ and $\mathcal{X}^+$ linearly separable with high probability? Is there a lower bound for the induced margin?*

Formally, this means that there exists a hyperplane $H[\boldsymbol{u}, \tau] \coloneqq \{\boldsymbol{z} \in \mathbb{R}^{\hat{n}} \mid \langle \boldsymbol{u}, \boldsymbol{z}\rangle + \tau = 0\}$ with $\|\boldsymbol{u}\|_2 = 1$ and $\tau \in \mathbb{R}$ that separates $F(\mathcal{X}^-)$ and $F(\mathcal{X}^+)$ with a certain *margin* $\mu > 0$:

$$\begin{aligned} \langle \boldsymbol{u}, F(\boldsymbol{x}^-)\rangle + \tau \leq -\mu \qquad &\text{for all } \boldsymbol{x}^- \in \mathcal{X}^-, \\ \langle \boldsymbol{u}, F(\boldsymbol{x}^+)\rangle + \tau \geq +\mu \qquad &\text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+. \end{aligned}$$

As such, Problem 1 states a purely geometric question on the separation capacity of random NNs (see Fig. 1 for an illustration), but it has also immediate consequences for associated learning tasks (e.g., see [2, Thm. 15.4], [3], [4]). We will address Problem 1 for random NNs with the following type of layers:

Definition 2. We call $\Phi\colon \mathbb{R}^{n_{\text{in}}} \to \mathbb{R}^{n_{\text{out}}}$ a *random ReLU-layer* with maximal bias $\lambda \geq 0$ if $\Phi(\boldsymbol{x}) = \sqrt{\frac{2}{n_{\text{out}}}} \cdot \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$, where the weight matrix $\boldsymbol{W} \in \mathbb{R}^{n_{\text{out}} \times n_{\text{in}}}$ has independent standard Gaussian entries and the bias vector $\boldsymbol{b}$ is uniformly distributed on $[-\lambda, \lambda]^{n_{\text{out}}}$; as the naming suggests, the element-wise activation function is the *rectified linear unit* (*ReLU*), i.e., $\mathrm{ReLU}(s) \coloneqq \max\{0, s\}$ for $s \in \mathbb{R}$.

II. Main Results

Before stating a formal solution to Problem 1, we need to introduce two geometric parameters: The *covering number* of a bounded subset $\mathcal{X} \subset \mathbb{R}^d$ at scale $r > 0$ is given by

$$\mathcal{N}(\mathcal{X}, r) \coloneqq \min\Big\{N \mid \exists \boldsymbol{c}_1, \dots, \boldsymbol{c}_N \in \mathbb{R}^d \colon \mathcal{X} \subset \bigcup_{j \in [N]} \mathbb{B}^d(\boldsymbol{c}_j, r)\Big\},$$

i.e., the smallest number of Euclidean balls of radius r required to cover $\mathcal{X}$; here $\mathbb{B}^d(\boldsymbol{c}_j, r) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x} - \boldsymbol{c}_j\|_2 \leq r\}$. Moreover, the *(Gaussian) mean width* of $\mathcal{X}$ is defined as

$$w(\mathcal{X}) \coloneqq \mathbb{E}_{\boldsymbol{g}}\Big[\sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{g}, \boldsymbol{x}\rangle\Big],$$

where $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ denotes a standard Gaussian random vector. Both $\mathcal{N}(\mathcal{X}, r)$ and $w(\mathcal{X})$ are natural complexity measures, which are well-established in convex geometry, high-dimensional probability, and signal processing, e.g., see [5], [6], [7], [8].

Theorem 3. *There exist absolute constants $c, c', C > 0$ such that the following holds. For $R \geq 1$, let $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{B}^d(\boldsymbol{0}, R)$ be two δ-separated sets and let $\lambda \geq e\delta$ be such that $\lambda \gtrsim R\sqrt{\log(\lambda/\delta)}$. We set $N^- \coloneqq \mathcal{N}(\mathcal{X}^-, c\delta^2/\lambda)$, $N^+ \coloneqq \mathcal{N}(\mathcal{X}^+, c\delta^2/\lambda)$, and $\mathcal{D} \coloneqq w^2(\mathcal{X}^-) + w^2(\mathcal{X}^+)$. Assume that $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ and $\hat{\Phi}\colon \mathbb{R}^n \to \mathbb{R}^{\hat{n}}$ are two (independent) random ReLU-layers with maximal biases $\lambda, \hat{\lambda} \geq 0$, respectively, such that*

$$n \gtrsim d + \big(\tfrac{\lambda}{\delta}\big)^8 \cdot \Big(\lambda^{-2} \cdot \mathcal{D} + \log(2N^-N^+/\eta)\Big) \tag{1}$$

and

$$\begin{aligned} &\hat{\lambda} \gtrsim \big(\tfrac{\lambda}{\delta}\big)^4 \cdot \big(\lambda\sqrt{\mathcal{D}} + \lambda^2\big), \\ &\hat{n} \gtrsim \big(\tfrac{\hat{\lambda}}{\lambda^2}\big) \cdot \exp\Big(C \cdot \big(\mathcal{D} + \lambda^2\big) \cdot \big(\tfrac{\lambda}{\delta}\big)^8 \cdot \log(\lambda/\delta)\Big) \cdot \log(N^-/\eta). \end{aligned} \tag{2}$$

Then, given the two-layer random neural network $F \coloneqq \hat{\Phi}(\Phi(\cdot))$, with probability at least $1 - \eta$, the sets $F(\mathcal{X}^-), F(\mathcal{X}^+) \subset \mathbb{B}^{\hat{n}}(\boldsymbol{0}, \hat{\lambda})$ are linearly separable with margin

$$c' \cdot \big(\tfrac{\lambda^4}{\hat{\lambda}}\big) \cdot \exp\Big(-C \cdot \big(\mathcal{D} + \lambda^2\big) \cdot \big(\tfrac{\lambda}{\delta}\big)^8 \cdot \log(\lambda/\delta)\Big).$$

The most important feature of the above result are the conditions (1) and (2), which determine the required widths of the two random layers in F. While the condition on the width of the first layer is rather mild, condition (2) involves an exponential factor. However, this term is actually governed by the complexity parameter $\mathcal{D}$, which can be substantially smaller than the ambient dimension d for structured data sets. Typical examples are data residing on a low-dimensional manifold or finite point clouds (implying a memorization task). Consequently, the curse of dimensionality can be overcome in such cases. On a larger scale, Theorem 3 presents an *instance-specific* solution to Problem 1, which relates the network size explicitly to the geometric complexity of the underlying classes $\mathcal{X}^-$ and $\mathcal{X}^+$.

Further refinements—taking into account the mutual arrangement between $\mathcal{X}^-$ and $\mathcal{X}^+$—are also possible and can be found in [1].

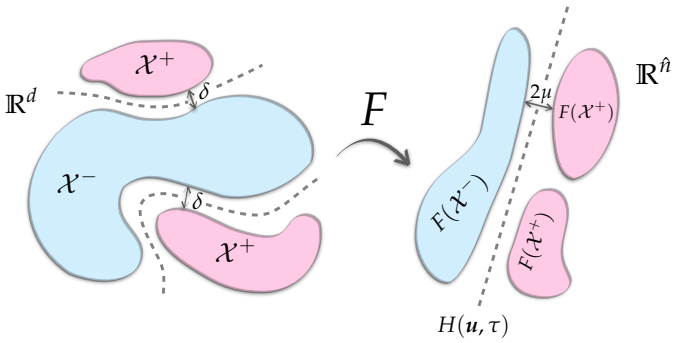

Fig. 1. **Illustration of Problem 1.** Can a random NN $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$ "disentangle" the two sets $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{R}^d$ such that they become linearly separable in the feature space $\mathbb{R}^{\hat{n}}$ with a positive margin μ? Except for being δ-separated and bounded, $\mathcal{X}^-$ and $\mathcal{X}^+$ may have an arbitrary decision boundary, possibly with multiple connected components.

REFERENCES

[1] S. Dirksen, M. Genzel, L. Jacques, and A. Stollenwerk, "The separation capacity of random neural networks," 2021, forthcoming.

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[3] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems 20*, 2007, pp. 1177–1184.

[4] ——, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Advances in Neural Information Processing Systems 21*, 2008, pp. 1313–1320.

[5] A. A. Giannopoulos and V. D. Milman, "Asymptotic convex geometry short overview," in *Different Faces of Geometry*, S. Donaldson, Y. Eliashberg, and M. Gromov, Eds. Springer Boston, 2004, pp. 87–162.

[6] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[7] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes*, ser. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer Berlin Heidelberg, 2014, vol. 3.

[8] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018, vol. 47.