Journal of Machine Learning Research 23 (2022) 1-47 Submitted 9/21; Revised 10/22; Published 11/22

The Separation Capacity of Random Neural Networks

Sjoerd Dirksen S.DIRKSEN@UU.NL
Mathematical Institute
Utrecht University
3584 CD Utrecht, Netherlands

Martin Genzel MARTINGENZEL@GMAIL.COM
Mathematical Institute
Utrecht University
3584 CD Utrecht, Netherlands

Laurent Jacques LAURENT.JACQUES@UCLOUVAIN.BE
ISPGroup, INMA, ICTEAM Institute
Université Catholique de Louvain
1348 Louvain-la-Neuve, Belgium

Alexander Stollenwerk ALEXANDER.STOLLENWERK@UCLOUVAIN.BE
ISPGroup, INMA, ICTEAM Institute
Université Catholique de Louvain
1348 Louvain-la-Neuve, Belgium

Editor: Joan Bruna

Abstract

Neural networks with random weights appear in a variety of machine learning applications, most prominently as the initialization of many deep learning algorithms and as a computationally cheap alternative to fully learned neural networks. In the present article, we enhance the theoretical understanding of random neural networks by addressing the following data separation problem: under what conditions can a random neural network make two classes $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{R}^d$ (with positive distance) linearly separable? We show that a sufficiently large 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 $\mathcal{X}^-, \mathcal{X}^+$ and their mutual arrangement. 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. We quantify the relevant structure of the data in terms of a novel notion of mutual complexity (based on a localized version of Gaussian mean width), which leads to sound and informative separation guarantees. We connect our result with related lines of work on approximation, memorization, and generalization.

Keywords: Random neural networks, classification, hyperplane separation, high-dimensional geometry, Gaussian mean width

1. Introduction

Despite the unprecedented success of neural networks (NNs) in countless applications (LeCun et al., 2015; Schmidhuber, 2015; Goodfellow et al., 2016), a rigorous understanding of

©2022 Sjoerd Dirksen, Martin Genzel, Laurent Jacques, and Alexander Stollenwerk.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-1079.html.

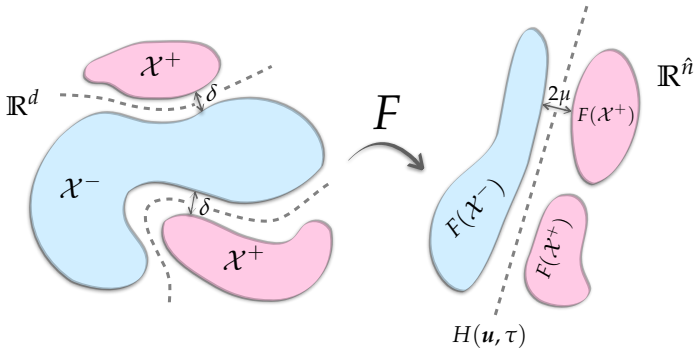

Figure 1: **Illustration of Problem 1.** Can a random NN $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$ "disentangle" the two sets $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{R}^d$ such that they become linearly separable in the feature space $\mathbb{R}^{\hat{n}}$ with a positive margin μ? Except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

their operating principles is still in its infancy. The present work is devoted to a mathematical study of *random NNs*, i.e., feedforward NNs whose weight parameters are drawn from a generic probability distribution. Random NNs play an important role in machine learning in at least three different ways. First, it is standard to initialize the training of a (deep) NN by random weights and it is well known that this initialization is a key contributor to the exceptional performance of NNs (He et al., 2015; Goodfellow et al., 2016; Arpit and Bengio, 2019). Second, it has been empirically observed that architecture search can be effectively carried out with random NNs, in the sense that the hierarchy in performance of fully trained architectures closely matches the hierarchy of the architectures with random weights (Saxe et al., 2011). Finally, random NNs have been extensively investigated as a cheap computational alternative to fully trained NNs: it has been demonstrated empirically that pre-processing with a random NN and applying a simple classification method already gives surprisingly good results (Huang et al., 2006; Rahimi and Recht, 2008; Zhang et al., 2017). For these reasons, it is of substantial interest to gain a deeper theoretical understanding of the properties of random NNs.

In this work, we shed new light on the capabilities of random NNs as a pre-processor by addressing the following fundamental problem on class separability:

Problem 1 *Consider two bounded, possibly infinite sets $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{R}^d$ that are* δ-separated*, i.e.,*

$$\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \geq \delta \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+ \text{ and } \boldsymbol{x}^- \in \mathcal{X}^-.$$

Let $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$ represent a (multi-layer) feedforward NN with random weights*, where the architecture 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 separation margin?*

Formally, we will identify conditions that ensure the existence (with high probability) of 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* $\mu > 0$, i.e.,

$$\begin{aligned} \langle \boldsymbol{u}, F(\boldsymbol{x}^-) \rangle + \tau \leq -\mu \qquad & \text{for all } \boldsymbol{x}^- \in \mathcal{X}^-, \\ \langle \boldsymbol{u}, F(\boldsymbol{x}^+) \rangle + \tau \geq +\mu \qquad & \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+. \end{aligned} \tag{1}$$

Problem 1 thus states a purely geometric question on the *separation capacity* of random NNs, see Figure 1 for an illustration. However, it is useful to bear in mind that the ability to render two "intertwined" sets linearly separable also has immediate consequences for associated learning tasks. To see this, let $(\boldsymbol{x}, y)$ be drawn from an arbitrary data distribution on $\mathbb{R}^d \times \{\pm 1\}$ satisfying

$$\mathbb{P}(\boldsymbol{x} \in \mathcal{X}^+ \mid y = +1) = 1 = \mathbb{P}(\boldsymbol{x} \in \mathcal{X}^- \mid y = -1),$$

i.e., the binary label Y is consistent with the classes $\mathcal{X}^-$ and $\mathcal{X}^+$. Conditioned on the high-probability event of Problem 1, the transformed pair $(F(\boldsymbol{x}), y)$ then fulfills a *hard-margin condition*:

$$\mathbb{P}\big(y \cdot (\langle \boldsymbol{u}, F(\boldsymbol{x}) \rangle + \tau) \geq \mu\big) = 1, \tag{2}$$

where $H[\boldsymbol{u}, \tau]$ denotes the separating hyperplane in (1). Given i.i.d. training samples of $(\boldsymbol{x}, y)$, this enables us to learn the unknown output parameters $(\boldsymbol{u}, \tau)$ by standard classification methods, such as support vector machines (SVMs) (Steinwart and Christmann, 2008). In particular, one can achieve provable control over the generalization error in terms of the margin size μ, e.g., see Shalev-Shwartz and Ben-David (2014, Thm. 15.4). Of similar relevance is the width $\hat{n}$ of the output-layer of F, as it determines the ambient dimension of the feature space and therefore the computational complexity of the classification method. For these reasons, we seek to solve Problem 1 with reasonable bounds for both μ and $\hat{n}$.

In principle, the aforementioned idea of using random NNs as a pre-processing step for well-understood (linear) classifiers is not new (Huang et al., 2006; Rahimi and Recht, 2008; Zhang et al., 2017). But despite conceptual overlaps, the analytical approach of the present article is different from most existing works, see also Section 1.5 for a short literature overview. Although Problem 1 includes a large family of classification tasks — namely all pairs of δ-separated sets — we are primarily interested in an *instance-specific* analysis: our main results quantify the dependence of the key parameters μ and $\hat{n}$ as functions of the underlying classes $\mathcal{X}^-$ and $\mathcal{X}^+$ and their "interaction". The resulting instance-specific bounds allow us to avoid pessimistic (worst-case) bounds caused by pathological pairs of sets. In our analysis, we make no explicit assumptions about the data in Problem 1, such as a handcrafted generative model or sampling from a generic distribution. Instead, we will introduce complexity measures that quantify the geometric complexities of $\mathcal{X}^-$ and $\mathcal{X}^+$ as well as their mutual entanglement. This perspective appears more natural to us in the context of data-driven methods.

Let us now specify the class of random NNs for which we will explore Problem 1. Throughout, the function $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$ will be composed of (hidden) layers of the following form.

Definition 1 *We call* $\Phi\colon \mathbb{R}^{n_{in}} \to \mathbb{R}^{n_{out}}$ *a* random ReLU-layer *with maximal bias* $\lambda \geq 0$ *if*

$$\Phi(\boldsymbol{x}) = \sqrt{\tfrac{2}{n_{out}}} \cdot \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \qquad \boldsymbol{x} \in \mathbb{R}^{n_{in}}, \tag{3}$$

where the weight matrix $\boldsymbol{W} \in \mathbb{R}^{n_{out}\times n_{in}}$ *has standard Gaussian entries, the bias vector* $\boldsymbol{b}$ *is uniformly distributed on* $[-\lambda,\lambda]^{n_{out}}$*, independently of* $\boldsymbol{W}$*, and the element-wise activation function is the rectified linear unit (ReLU), i.e.,* $\mathrm{ReLU}(s) \coloneqq \max\{0, s\}$ *for* $s \in \mathbb{R}$*.*

It will turn out that already two random ReLU-layers are sufficient for our solution to Problem 1, although in principle deeper architectures are also possible. We consider the ReLU mainly because of its popularity, but our analysis can be adapted for other common activation functions, e.g., the thresholding activation. Let us note that the random weights and normalization in Definition 1 do not exactly correspond to a standard initialization in deep learning. The closest is the popular *He initialization* (He et al., 2015), which would be obtained by replacing n_{out} by n_{in} and taking $\boldsymbol{b} = \boldsymbol{0}$ in (3). Our non-standard choice of the bias is due to a hyperplane tessellation argument used in the proof of our main result. The proof sketch in Section 1.4 will provide an intuitive explanation for this choice; in particular, see Figure 3(b).

Instead of directly formulating our main result, Theorem 10, we will first present several readily accessible special cases of increasing generality. Afterwards, we will highlight our proof strategy in Section 1.4.

1.1 A Gentle Start: Finite Sets and Memorization

Our first result below gives an answer to Problem 1 in the situation of *finite* point sets. In the following, $\mathbb{B}_2^d \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 \leq 1\}$ denotes the Euclidean unit ball; see also Section 1.6 for a summary of common notation used in this article.

Theorem 2 (Finite sets) *There exist absolute constants* $c, C > 0$ *such that the following holds.*

Let $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{B}_2^d$ *be* δ*-separated sets with* $N^+ \coloneqq |\mathcal{X}^+| < \infty$ *and* $N^- \coloneqq |\mathcal{X}^-| < \infty$*. Suppose that* $\lambda \gtrsim \sqrt{\log(e\lambda/\delta)}$*. Let* $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ *and* $\hat{\Phi}\colon \mathbb{R}^n \to \mathbb{R}^{\hat{n}}$ *be two (independent) random ReLU-layers with maximal biases* $\lambda, \hat{\lambda} \geq 0$*, respectively, such that*

$$n \gtrsim \left(\tfrac{\lambda}{\delta}\right)^8 \cdot \log(2N^-N^+/\eta) \tag{4}$$

and

$$\hat{\lambda} \gtrsim \left(\tfrac{\lambda}{\delta}\right)^4 \cdot \big(\alpha + \lambda\big), \qquad \hat{n} \gtrsim \tfrac{\hat{\lambda}}{\lambda} \cdot \theta \cdot \log(N^-/\eta), \tag{5}$$

where $\alpha = \sqrt{\log N^+}$ *and*

$$\theta = \exp\Big(C \cdot \big(\alpha^2 + \lambda^2\big) \cdot \lambda^6 \cdot \delta^{-8} \cdot \log(\lambda/\delta)\Big). \tag{6}$$

Then, given the two-layer random NN $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$*,* $\boldsymbol{x} \mapsto \hat{\Phi}(\Phi(\boldsymbol{x}))$*, with probability at least* $1-\eta$*, the sets* $F(\mathcal{X}^-), F(\mathcal{X}^+) \subset \hat{\lambda}\mathbb{B}_2^{\hat{n}}$ *are linearly separable with margin* $c\lambda^2/(\hat{\lambda}\theta)$*.*

In this result, the best choices for λ and $\hat{\lambda}$ are the minimal settings that satisfy the stated bounds. The governing condition in Theorem 2 (and in all following results, Theorems 3, 4, and 10 below) is condition (5) on the width $\hat{n}$ of the second layer. It features the term θ that scales exponentially in terms of the logarithm of the number of points, so that $\hat{n}$ scales as $N_+^{\mathrm{poly}(\lambda,1/\delta)}$, in contrast to the logarithmic scaling of n in (4). To gauge whether this condition is necessary, let us connect Theorem 2 to the *memorization capacity* of random NNs. The ability of memorizing large data sets (including their noisy components) is a well-known phenomenon in deep learning research and considered as an important piece of the still unsolved generalization puzzle (Zhang et al., 2017, 2021). Theorem 2 applies to any (δ-separated) completely unstructured data set — imagine a point cloud with arbitrary binary labels. Remarkably, one can therefore memorize the labels of any such (finite) set with high probability by efficiently computing a separating hyperplane of $F(\mathcal{X}^-)$ and $F(\mathcal{X}^+)$, e.g., using a hard-margin SVM.[1]

Although this shows that random NNs can be powerful memorizers in practice, existing results in the literature indicate that perfect memorization is already possible when the number of neurons scales linearly in $(N^- + N^+)$ up to logarithmic factors, e.g., see Yun et al. (2019); Vershynin (2020); Bresler and Nagaraj (2020). Hence, we expect that the dependence on δ and λ within the exponential term θ in (5) may be improved.

1.2 Separation of Euclidean Balls

Although Theorem 2 provides a margin bound, its actual size was not relevant to the network's memorization capacity. The situation is different for *infinite classes*, on which we will focus from now on. Problem 1 is then connected to a binary classification task through the hard-margin condition (2), and the margin size determines the generalization performance (Shalev-Shwartz and Ben-David, 2014, Thm. 15.4). Our next result may be seen as a natural extension of Theorem 2, replacing discrete data points by a finite collection of Euclidean balls; see Figure 2 for an illustration of this model.

Theorem 3 (Euclidean balls) *There exist absolute constants $c, C > 0$ such that the following holds.*

Let $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{B}_2^d$ be δ-separated sets that can be written as the union of finitely many Euclidean balls of radius $r \geq 0$, i.e.,

$$\mathcal{X}^- = \bigcup_{l\in[N^-]} \mathbb{B}_2^d(\boldsymbol{c}_l^-, r), \quad \mathcal{X}^+ = \bigcup_{j\in[N^+]} \mathbb{B}_2^d(\boldsymbol{c}_j^+, r).$$

Suppose that $\lambda \gtrsim \sqrt{\log(e\lambda/\delta)}$ and $r \lesssim \delta^2/\lambda$. We 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

$$n \gtrsim (1 + \lambda^6\delta^{-8}r^2) \cdot d + \left(\tfrac{\lambda}{\delta}\right)^8 \cdot \log(2N^-N^+/\eta) \tag{7}$$

and (5) holds with $\alpha = r\sqrt{d} + \sqrt{\log N^+}$ and θ as in (6). Then, given the two-layer random NN $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$, $\boldsymbol{x} \mapsto \hat{\Phi}(\Phi(\boldsymbol{x}))$, with probability at least $1-\eta$, the sets $F(\mathcal{X}^-), F(\mathcal{X}^+) \subset \hat{\lambda}\mathbb{B}_2^{\hat{n}}$ are linearly separable with margin $c\lambda^2/(\hat{\lambda}\theta)$.

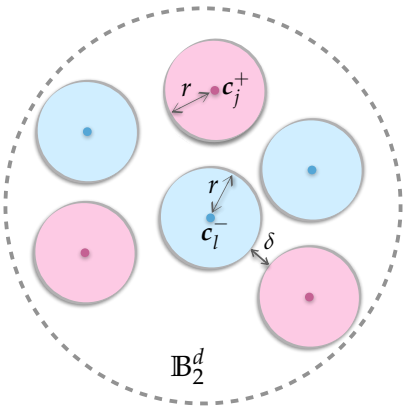

Figure 2: **Illustration of the Euclidean-ball model in Theorem 3.** The sets $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{B}_2^d$ consist of Euclidean balls of radius r, where the center points are denoted by $\boldsymbol{c}_l^-$ and $\boldsymbol{c}_j^+$, respectively. Note that the δ-separation only concerns balls of different classes, while arbitrary intersections are allowed within each class.

It is worth noting that in the limit case $r = 0$, the above statement is essentially consistent with Theorem 2. On the other hand, Theorem 3 reveals the price of dealing with full-dimensional sets instead of points: the bound on the output dimension $\hat{n}$ scales exponentially in terms of $r^2 d$. Thus, to avoid the curse of dimensionality, the radius needs to satisfy $r \lesssim 1/\sqrt{d}$. Under this assumption, Theorem 3 certifies that random NNs can efficiently separate (unstructured) collections of Euclidean balls.

1.3 Towards a General Separation Guarantee

So far, we have only considered specific examples of data sets. Our next theorem concerns Problem 1 for arbitrary δ-separated classes. For a formal statement, we need to introduce two important geometric parameters. The *covering number* of a bounded subset $\mathcal{X} \subset \mathbb{R}^d$ at scale $r > 0$ is given by

$$\mathcal{N}(\mathcal{X}, r) \coloneqq \min\left\{N \in \mathbb{N} \mid \exists \boldsymbol{c}_1, \dots, \boldsymbol{c}_N \in \mathbb{R}^d \colon \mathcal{X} \subset \bigcup\nolimits_{j\in[N]} \mathbb{B}_2^d(\boldsymbol{c}_j, r)\right\}, \tag{8}$$

i.e., the smallest number of Euclidean balls of radius r required to cover $\mathcal{X}$. 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 high-dimensional geometry,

1. This observation is especially interesting when no obvious learning rule is available. On the other hand, if the data carries more structure (e.g., $\mathcal{X}^-$ and $\mathcal{X}^+$ are already linearly separable), there certainly exist more effective approaches than randomized transforms.

statistics, and signal processing, e.g., see Giannopoulos and Milman (2004); Chandrasekaran et al. (2012); Talagrand (2014); Vershynin (2018).

Theorem 4 (General sets) *There exist absolute constants $c, C > 0$ such that the following holds.*

Let $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{B}_2^d$ be δ-separated sets and suppose that $\lambda \gtrsim \sqrt{\log(e\lambda/\delta)}$. Moreover, let $N^- := \mathcal{N}(\mathcal{X}^-, c\delta^2/\lambda)$ and $N^+ := \mathcal{N}(\mathcal{X}^+, c\delta^2/\lambda)$. We 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[2]

$$\begin{aligned} n &\gtrsim w^2(\mathrm{cone}(\mathcal{X}^- - \mathcal{X}^-) \cap \mathbb{S}^{d-1}) + w^2(\mathrm{cone}(\mathcal{X}^+ - \mathcal{X}^+) \cap \mathbb{S}^{d-1}), \\ n &\gtrsim \left(\tfrac{\lambda}{\delta}\right)^8 \cdot \Big(\lambda^{-2}\big(w^2(\mathcal{X}^-) + w^2(\mathcal{X}^+)\big) + \log(2N^-N^+/\eta)\Big) \end{aligned} \tag{9}$$

and (5) holds with $\alpha = w(\mathcal{X}^-) + w(\mathcal{X}^+)$ and θ as in (6). Then, given the two-layer random NN $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$, $\boldsymbol{x} \mapsto \hat{\Phi}(\Phi(\boldsymbol{x}))$, with probability at least $1 - \eta$, the sets $F(\mathcal{X}^-), F(\mathcal{X}^+) \subset \hat{\lambda}\mathbb{B}_2^{\hat{n}}$ are linearly separable with margin $c\lambda^2/(\hat{\lambda}\theta)$.

Compared to Theorem 3, the above guarantee yields a much stronger statement due to the use of the mean width as a complexity measure. To see this, let $\mathcal{X}^+ \subset \bigcup_{j\in[N^+]} \mathbb{B}_2^d(\boldsymbol{c}_j^+, r)$ be any covering of $\mathcal{X}^+$ at scale $r = c\lambda^{-1}\delta^2$ and consider the following upper bound (see Lemma 29):

$$w(\mathcal{X}^+) \lesssim w^+ + \sqrt{\log N^+}, \tag{10}$$

where $w^+ = \max_{j\in[N^+]} w(\mathcal{X}^+ \cap \mathbb{B}_2^d(\boldsymbol{c}_j^+, r))$. An analogous bound holds for $\mathcal{X}^-$. While using the worst-case estimate $w^+ \lesssim r\sqrt{d}$ would lead to a similar bottleneck as in Theorem 3, the *localized* mean width parameter w^+ can be substantially smaller for structured data sets. Typical examples are data residing on a low-dimensional manifold or contained in the convex hull of finitely many points, see Remark 5 below for some concrete examples. On the other hand, the covering number N^+ reflects the *global* size of $\mathcal{X}^+$ in (10).

For these reasons, Theorem 4 takes an important step towards a general solution to Problem 1, which meets our overall goal of *instance-specific* bounds for the separation margin and layer widths. Having said this, the geometric parameters in this result only capture the individual complexities of $\mathcal{X}^-$ and $\mathcal{X}^+$, but remain silent about their mutual arrangement. For instance, one would expect that two sets become easier to separate if their "centers of mass" are farther apart, even though the minimal distance δ is small; see Figure 5 in Section 2 for an illustration. In such scenarios, a non-uniform covering strategy for $\mathcal{X}^-$ and $\mathcal{X}^+$ is preferable, in the sense that data points far away from the decision boundary should be covered by fewer but larger balls. The most general outcome of this work, Theorem 10, makes this intuition precise by employing a novel notion of *mutual complexity* (see Definition 8 and 9). We refer to Section 2 for an in-depth discussion and further refinements due to Theorem 10. Finally, we emphasize that all previously presented results follow from Theorem 10 as special cases, see Section 6 for detailed proofs.

We close this part with a few examples of concrete bounds on the mean width to highlight its usefulness as a complexity measure:

2. Note that the first line in (9) is always satisfied if $n \gtrsim d$, since we have that $w^2(\mathbb{S}^{d-1}) \asymp d$. See also Section 1.6 for a precise definition of $\mathrm{cone}(\cdot)$.

Remark 5 (Controlling the mean width) *(1)* Worst-case bound. *Since $\mathcal{X}^-, \mathcal{X}^+ \subset \mathbb{B}_2^d$, the mean width parameter in Theorem 4 satisfies the trivial bound*

$$\alpha = w(\mathcal{X}^-) + w(\mathcal{X}^+) \lesssim w(\mathbb{B}_2^d) \asymp \sqrt{d}.$$

Thus, an exponential width of the second layer in terms of d allows us to solve Problem 1 for arbitrary δ-separated sets, regardless of their specific shape.

(2) Low-dimensional subspaces. *As highlighted above, already much smaller networks can achieve separation if the underlying classes carry more structure. A typical example of low-complexity structure is a situation where $\mathcal{X}^-$ and $\mathcal{X}^+$ reside in a union of low-dimensional subspaces, say $\mathcal{X}^-, \mathcal{X}^+ \subset \bigcup_{j\in[N]} L_j \cap \mathbb{B}_2^d$ with $\dim L_j \ll d$. Then,*

$$\alpha = w(\mathcal{X}^-) + w(\mathcal{X}^+) \lesssim \max_{j\in[N]} \sqrt{\dim L_j} + \sqrt{\log N} \ll \sqrt{d},$$

assuming that N is not exponentially large (see Lemma 29).

(3) Point clouds and their convex hulls. *Another important example of a low-complexity set is the convex hull of finitely many points. Indeed, assuming $\mathcal{X}^-, \mathcal{X}^+ \subset \operatorname{conv}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) \subset \mathbb{B}_2^d$, the mean width only scales logarithmically in N:*

$$\alpha = w(\mathcal{X}^-) + w(\mathcal{X}^+) \le 2 \cdot w(\operatorname{conv}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)) = 2 \cdot w(\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}) \lesssim \sqrt{\log N},$$

where we have used a basic bound on the mean width (e.g., see Vershynin, 2015, Ex. 1.3.8) and its invariance under taking the convex hull. Note that this bound particularly extends the situation of finitely many data points from Theorem 2 to infinite data sets.

1.4 Proof Strategy

To keep our exposition as simple as possible, we will describe our proof strategy in the prototypical situation of Euclidean balls from Theorem 3, see also Figure 2. Recall that $\mathcal{C}^- := \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\}$ and $\mathcal{C}^+ := \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\}$ denote the center points of $\mathcal{X}^-$ and $\mathcal{X}^+$, respectively. For convenience, we also set $\mathcal{X}_l^- := \mathbb{B}_2^d(\boldsymbol{c}_l^-, r)$ and $\mathcal{X}_j^+ := \mathbb{B}_2^d(\boldsymbol{c}_j^+, r)$ so that $\mathcal{X}^- = \bigcup_{l\in[N^-]} \mathcal{X}_l^-$ and $\mathcal{X}^+ = \bigcup_{j\in[N^+]} \mathcal{X}_j^+$.

Our data separation approach consists of a *two-step procedure*, which essentially corresponds to the composition of the random layers $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ and $\hat{\Phi}\colon \mathbb{R}^n \to \mathbb{R}^{\hat{n}}$. Although Φ and $\hat{\Phi}$ both follow exactly the same random design (see Definition 1), we will see that their purposes are different: while the first one already establishes a desirable geometrical configuration under mild conditions, a major challenge is to show that one can actually take advantage of it by applying a second (wider) random layer. The central finding of our geometric analysis of Problem 1 is a subtle interplay between the separation capacity of random NNs and their stability properties. In particular, we demonstrate that the linearization of complicated data is possible on a global scale, without too much disturbing its local geometry, e.g., Euclidean point distances.

To understand the effect of the first layer Φ, it is useful to take a coordinate-wise perspective:

$$[\Phi(\boldsymbol{x})]_i = \sqrt{\tfrac{2}{n}}\,\mathrm{ReLU}(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle + b_i), \quad i = 1, \dots, n, \quad \boldsymbol{x} \in \mathbb{R}^d,$$

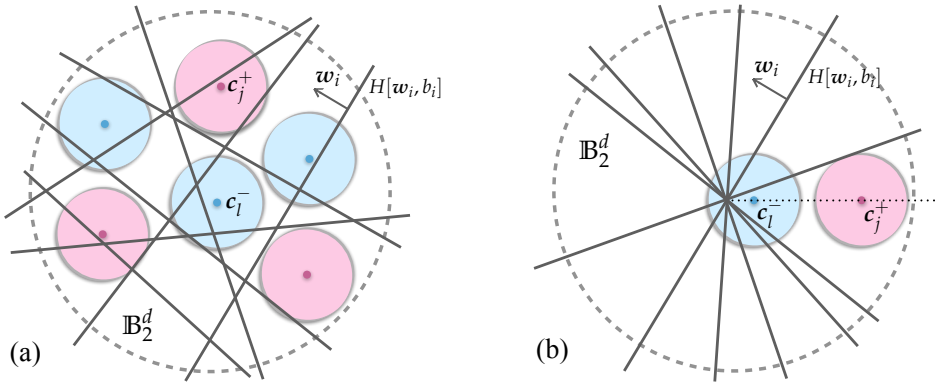

Figure 3: **Random hyperplanes in the input domain.** (a) Each coordinate of the first layer $\Phi : \mathbb{R}^d \to \mathbb{R}^n$ can be associated with a random hyperplane $H[\boldsymbol{w}_i, b_i]$, where $\boldsymbol{w}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and b_i is uniformly distributed on $[-\lambda, \lambda]$. Already a relatively few of such hyperplanes are enough to separate every pair of center points $(\boldsymbol{c}_l^-, \boldsymbol{c}_j^+) \in \mathcal{C}^- \times \mathcal{C}^+$ at least once. For this, a sufficiently large bias parameter ($\lambda \gtrsim 1$) is vital, as it ensures a *uniform tessellation* of the input domain $\mathbb{B}_2^d$; otherwise, the probability of separating points close to the boundary of $\mathbb{B}_2^d$ would become too low. Subfigure (b) illustrates what could go wrong for $\lambda = 0$: if the center points reside on a ray starting at the origin, a separation by hyperplanes without offsets becomes impossible.

where $\boldsymbol{w}_i \in \mathbb{R}^d$ is the i-th row of the weight matrix $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ and $b_i \in [-\lambda, \lambda]$ the corresponding bias. Thus, $[\Phi(\boldsymbol{x})]_i$ indicates on which side of the (unnormalized) random hyperplane $H[\boldsymbol{w}_i, b_i]$ a given point $\boldsymbol{x} \in \mathbb{R}^d$ lies. Our first major proof step (elaborated in Theorem 26) shows that as long as $\lambda \gtrsim 1$, the following holds with high probability: for every pair of center points $(\boldsymbol{c}_l^-, \boldsymbol{c}_j^+) \in \mathcal{C}^- \times \mathcal{C}^+$, there are coordinates $I_{l,j} \subset [n]$ with $|I_{l,j}| \gtrsim \delta \lambda^{-1} n$ such that $H[\boldsymbol{w}_i, b_i]$ separates $\boldsymbol{c}_l^-$ from $\boldsymbol{c}_j^+$ for all $i \in I_{l,j}$, in fact

$$[\Phi(\boldsymbol{c}_l^-)]_i = 0 \quad \text{and} \quad [\Phi(\boldsymbol{c}_j^+)]_i \gtrsim \tfrac{\delta}{\sqrt{n}}, \tag{11}$$

see Figure 3 for an illustration. The key insight to show (11) is that the probability of a single random hyperplane separating a fixed pair of δ-separated points is of order $\Omega(\delta \lambda^{-1})$, see Theorem 18. Combining this with a Chernoff bound over all hyperplanes associated with Φ then leads to a high-probability event of the above type. A remarkable fact about this argument is that the required layer width n is very moderate, scaling only logarithmically with the number of center points (see (7)).

A crucial part of (11) is that the corresponding coordinates in $\Phi(\boldsymbol{c}_l^-)$ are actually vanishing, due to the *non-linear* activation. Based on this, we can explicitly construct a (normalized) hyperplane $H[\boldsymbol{u}_l, \tau_l]$ for each $l \in [N^-]$ that separates $\Phi(\boldsymbol{c}_l^-)$ and $\Phi(\mathcal{C}^+)$ with margin $\tilde{\mu} \asymp \delta^2 \lambda^{-1}$ (see Theorem 26). The resulting arrangement of the transformed sets

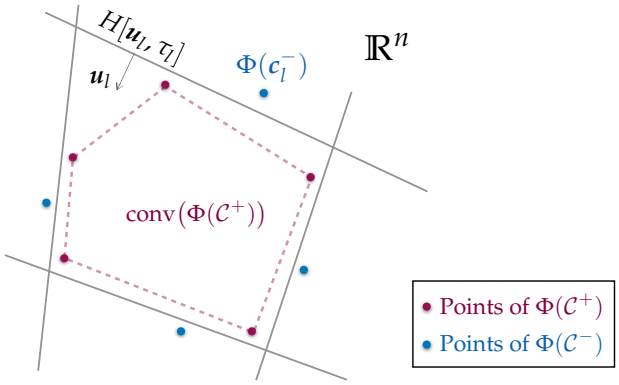

Figure 4: **The geometric effect of the first random ReLU-layer.** For each $l \in [N^-]$, the sets $\Phi(\boldsymbol{c}_l^-)$ and $\Phi(\mathcal{C}^+)$ are linearly separable, or equivalently, it holds that $\Phi(\mathcal{C}^-) \cap \mathrm{conv}(\Phi(\mathcal{C}^+)) = \emptyset$, where $\mathrm{conv}(\cdot)$ is the convex hull operator. One can picture $\mathrm{conv}(\Phi(\mathcal{C}^+))$ as a big "planet" which is orbited by small "satellites" namely the transformed center points in $\Phi(\mathcal{C}^-)$. For symmetry reasons, an analogous statement holds with high probability if the roles of $\mathcal{C}^-$ and $\mathcal{C}^+$ are interchanged.

$\Phi(\mathcal{C}^-)$ and $\Phi(\mathcal{C}^+)$ resembles a big "planet" which is orbited by small "satellites" and is illustrated in Figure 4.[3]

To conclude with the first layer, we need to ensure that the simplification achieved by Φ does not only apply to the center points but to the entire data set. Indeed, leveraging the *geometry-preserving* properties of random ReLU-layers (i.e., Φ preserves ℓ^2-distances between nearby points; see Theorem 19), it follows that $\Phi(\mathcal{X}_l^-)$ and $\Phi(\mathcal{X}^+)$ are also linearly separable for every $l \in [N^-]$ (still with margin $\tilde{\mu} \asymp \delta^2 \lambda^{-1}$, see Theorem 27). Hence, the geometric picture of Figure 4 remains true when replacing $\mathcal{C}^-$ and $\mathcal{C}^+$ by $\mathcal{X}^-$ and $\mathcal{X}^+$, respectively.[4] As for (11), this reasoning requires the bias $\boldsymbol{b}$ to be large enough (i.e., $\lambda \gtrsim \sqrt{\log(\lambda/\delta)}$) and exploits that it is *uniformly* distributed.

Let us now turn to the second layer $\hat{\Phi} \colon \mathbb{R}^n \to \mathbb{R}^{\hat{n}}$, which builds directly on the geometric situation after applying Φ. It is again helpful to treat each coordinate individually:

$$[\hat{\Phi}(\boldsymbol{x})]_i = \sqrt{\tfrac{2}{\hat{n}}}\,\mathrm{ReLU}(\langle \hat{\boldsymbol{w}}_i, \boldsymbol{x} \rangle + \hat{b}_i), \quad i = 1, \dots, \hat{n}, \quad \boldsymbol{x} \in \mathbb{R}^n,$$

3. Note that such a geometric arrangement would not be achievable without some kind of non-linearity in Φ. For example, imagine a series of points on a straight line where the class label ± 1 alternates with each point. This arrangement cannot be transformed into the situation in Figure 4 by an affine map, which maps lines to lines.
4. Using such distance preservation properties of Φ is very different from a direct approach, according to which a random hyperplane $H[\boldsymbol{w}_i, b_i]$ would separate pairs of balls $(\mathcal{X}_l^-, \mathcal{X}_j^+)$. In fact, the latter event is much less likely than the separation of single points by a random hyperplane (this will become clear with Theorem 6 below). Instead, distance preservation is an integral property of the random layer, which exploits information from *all* coordinates of Φ.

where $\hat{\boldsymbol{w}}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and $\hat{b}_i$ is uniformly distributed on $[-\hat{\lambda}, \hat{\lambda}]$. Our main goal is to show that for every $l \in [N^-]$ there exist sufficiently many coordinates $i \in [\hat{n}]$ such that

$$[\hat{\Phi}(\Phi(\mathcal{X}_l^-))]_i \geq \tfrac{t}{\sqrt{\hat{n}}} \quad \text{and} \quad [\hat{\Phi}(\Phi(\mathcal{X}^+))]_i = 0, \tag{12}$$

where $t > 0$ depends on the complexity of $\mathcal{X}^-$ and $\mathcal{X}^+$; note that the vanishing coordinates are associated with data from $\mathcal{X}^+$ instead of $\mathcal{X}^-$. Our basic strategy to establish (12) is similar to (11), but there is a major difference: we now have to deal with the probability that a random hyperplane separates the sets $\hat{\mathcal{X}}_l^- \coloneqq \Phi(\mathcal{X}_l^-)$ and $\hat{\mathcal{X}}^+ \coloneqq \Phi(\mathcal{X}^+)$. The outcome of the first layer implies the existence of a separator, e.g., $H[\boldsymbol{u}_l, \tau_l]$ (see also Figure 4), but this does not mean that it is likely to be found by a single random draw.[5] Certainly, the probability of successful separation may not only depend on the distance between $\hat{\mathcal{X}}_l^-$ and $\hat{\mathcal{X}}^+$, but also on their complexity and mutual arrangement. The following result makes this concern precise and forms a key component of our analysis; its proof can be found in Section 3.1. The notion of (ε, γ)-linear separability used below is formally introduced in Definition 12; for now, it is useful to think of a refinement of linear separability that captures how much a separating hyperplane for two sets can be perturbed, such that it is still separates the sets.

Theorem 6 *There exists an absolute constant $C > 0$ such that the following holds.*

For $\varepsilon \in [0, 1]$, $\gamma > 0$, and $R \geq 1$, let $\mathcal{X}^-, \mathcal{X}^+ \subset R\mathbb{B}_2^d$ be two (ε, γ)-linearly separable sets[6] *and put $\mu \coloneqq (1 - \varepsilon)\gamma$. Let $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ be a standard Gaussian vector and let τ be uniformly distributed on $[-\lambda, \lambda]$ for some $\lambda > 0$. For any $t \gtrsim w(\mathcal{X}^+ - \mathcal{X}^-) + R$ with $\lambda \gtrsim Rt\mu^{-1}$, the hyperplane $H[\boldsymbol{g}, \tau]$ t-separates $\mathcal{X}^-$ from $\mathcal{X}^+$ with probability at least*

$$\tfrac{t}{\lambda} \cdot \exp\big(-C \cdot t^2\mu^{-2} \cdot \log(4(1-\varepsilon)^{-1})\big). \tag{13}$$

While the factor t/λ may become small due to the condition $\lambda \gtrsim Rt\mu^{-1}$, the dominating term in (13) is the exponential one. In fact, the central element of Theorem 6 is the mean width $w(\mathcal{X}^+ - \mathcal{X}^-)$, which dictates the severity of the exponential decay in (13).

An appropriate combination of Theorem 6 with a Chernoff bound will allow us to derive a statement of the form (12). With this at hand, it is then relatively straightforward to show that $F(\mathcal{X}^-) = \hat{\Phi}(\Phi(\mathcal{X}^-))$ and $F(\mathcal{X}^+) = \hat{\Phi}(\Phi(\mathcal{X}^+))$ are indeed linearly separable (see Corollary 25). Noteworthy is that the resulting margin and the number of required neurons $\hat{n}$ both inherit the exponential scaling from (13), which is reflected in all presented separation guarantees. This observation particularly explains why the mean width appears as a natural measure of complexity for the data sets.

To the best of our knowledge, Theorem 6 is a new result and could be of independent interest: it concerns the fundamental question of when pairs of sets are likely to be separated by a random hyperplane and when not. Perhaps not very surprisingly, the probability of success might scale poorly in the worst case, which is an inevitable consequence of the concentration of measure phenomenon. But for highly structured (low-dimensional) sets, the situation can be much more benign; finite point sets as considered in our analysis of the first layer are a good example.

5. To be clear about this point, $H[\boldsymbol{u}_l, \tau_l]$ does explicitly depend on the unknown sets $\mathcal{X}^-$ and $\mathcal{X}^+$. Hence, in contrast to random hyperplanes, it cannot be used for practical purposes offhand.
6. Note that the notation for $\mathcal{X}^-$ and $\mathcal{X}^+$ is generic here. In the proof of our main result, we will apply Theorem 6 with $\mathcal{X}^- \coloneqq \hat{\mathcal{X}}^+$ and $\mathcal{X}^+ \coloneqq \hat{\mathcal{X}}_l^-$.

Remark 7 *Using the definition of (ε, γ)-linear separability (see Definition 12), one can show that $w(\mathcal{X}^+ - \mathcal{X}^-) \lesssim R\sqrt{\varepsilon d}$ holds in the setup of Theorem 6. This general upper bound indicates that in the worst case, the probability of separation may decrease exponentially with the ambient dimension d (unless $\varepsilon \lesssim \frac{1}{d}$).*

1.5 Related Literature

The two-step separation procedure underlying our proofs (see Section 1.4) is inspired by a construction of An et al. (2015). Their main result verifies that any two disjoint sets $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$ can be made linearly separable by a *deterministic* two-layer NN. However, An et al. (2015) show a pure existence statement and their method is not feasible from an algorithmic perspective, since the selected weight parameters explicitly depend on the sets to be separated; furthermore, no informative bounds for the number of required neurons are provided. By using random weights and suitable notions of complexity (namely mutual covering), we are able to derive much more practical separation guarantees, which eliminate the aforementioned shortcomings. This achievement entails novel mathematical ingredients, most notably the separation capacity of random hyperplanes (see Theorem 6) and uniform distance preservation by random ReLU-layers (see Theorem 19).

Below we will survey some works from the rich literature on random NNs that have notable conceptual similarities to our work. We are not aware of a comparable result that addresses the separation capacity of random NNs.

Approximation theory. A very active line of research investigates to what extent random NNs are *universal approximators* (e.g., see Andoni et al., 2014; Sun et al., 2018; Yehudai and Shamir, 2019; Needell et al., 2020; Hsu et al., 2021 and the references therein). Specifically, one considers a class of real-valued functions on a domain in $\mathbb{R}^d$ (e.g., continuous or Lipschitz functions), an approximation metric (typically the L^2- or L^∞-norm), and a shallow NN consisting of a ReLU-layer with random weights followed by a linear layer with arbitrary weights (that may depend on the function to be approximated). The aforementioned works quantify which size of the random layer guarantees that the NN can reach a pre-specified approximation error for every function in the given class. These results feature an exponential bottleneck, in the sense that the width of the random layer needs to scale exponentially in terms of the data dimension d to ensure accurate approximation (see also Needell et al. (2020) for a refinement if the domain is a lower-dimensional smooth manifold).

In principle, one could try to approach Problem 1 by applying such an approximation result to a function $f : \mathbb{R}^d \to \mathbb{R}$ that takes values $+1$ and -1 on $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively. However, existing approximation guarantees cannot certify a *zero* approximation error on the two sets. In addition, even if an existing result would apply, it would lead to a more pessimistic statement that involves a random ReLU-layer whose width scales exponentially in terms of d, rather than the more refined, instance-specific complexity measures considered here.

Learning with random features. The concept of random features was introduced by Rahimi and Recht (2007) as a cheap computational alternative to kernel methods. The idea is to construct a random feature map such that inner products between random data features approximate kernel evaluations of the original data, provided that the feature dimension

is high enough. A prime example are random Fourier features, which are designed to approximate the Gaussian kernel (Rahimi and Recht, 2007). Instead of a computationally expensive kernel method (e.g., kernel SVM), one can use a linear method (e.g., SVM) on the random features. A previous line of research has analyzed the generalization error of such methods, thereby quantifying the feature dimension that guarantees a performance on par with the associated kernel method (e.g., see Rahimi and Recht, 2008; Rudi and Rosasco, 2017; Bach, 2017; Sun et al., 2018; Li et al., 2021 and Liu et al., 2020 for a survey). Several of these results particularly apply when the feature map is a random ReLU-layer. Although these works indicate that the data is transformed in a beneficial way for learning, they do not have a direct connection to Problem 1. Perhaps the closest connection can be found in Cao and Gu, 2019b, where it is shown that if the random ReLU feature function class from Rahimi and Recht, 2008 can separate a finite set of data on the sphere, then a sufficiently wide random ReLU layer (without bias) can make the same data linearly separable with high probability, see Cao and Gu, 2019b, Asm. 4.10 and Lem. B.2. It is, however, unclear how to extend this statement to infinite datasets and how this separability assumption relates to the Euclidean separability assumption in Problem 1.

Neural tangent kernels and mean field regime. An intriguing finding of deep learning theory is that training randomly initialized NNs via gradient descent in the infinite-width limit is equivalent to kernel gradient descent with a specific type of kernel, called the neural tangent kernel (NTK); see Jacot et al., 2018. The behaviour in the infinite-width limit has partially motivated a line of work on the analysis of (stochastic) gradient descent for training NNs in the overparametrized regime, starting from a random initialization, e.g., see Arora et al., 2019; Oymak and Soltanolkotabi, 2019; Li and Liang, 2018; Du et al., 2019; Allen-Zhu et al., 2019; Zou and Gu, 2019; Cao and Gu, 2019a and the references therein. These works have roughly shown that (S)GD can achieve an arbitrarily small training (and sometimes even generalization) error if the NN is wide enough and, moreover, the (S)GD iterates remain close to the initialization. The required width of the NN is implicitly or explicitly linked to the NTK. Most closely connected to our work are Nitanda et al., 2019; Ji and Telgarsky, 2020; Chen et al., 2019, which explicitly link the required width to the separation capacity of the infinite-width NTK-feature map at initialization. As part of the analysis it is shown that if the training data satisfies a separability condition in the reproducing kernel Hilbert space induced by the infinite-width NTK, then the NTK-feature map associated with the finite-width random NN at initialization makes the training data linearly separable with high probability, e.g., see Ji and Telgarsky, 2020, Asm. 2.1, Lem. 2.3 & Sec. 5. These results bear resemblance with Problem 1, but there are several important differences. While we are primarily interested in the separation of *infinite* data sets, these works focus on finite-sample scenarios. It is not clear how the latter could be extended accordingly. Moreover, note that the NTK-feature map associated with a finite-width random NN is not a random NN itself.[7] Therefore, the aforementioned results do not address Problem 1 as such and the bounds on the network width needed to achieve linear separability are not directly comparable to ours.

7. If $F_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ is a NN with (randomly initialized) weights $\boldsymbol{\theta}$, then the associated NTK-feature map is $\boldsymbol{x} \mapsto \frac{\partial F_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}}$.

Let us mention for completeness that the connection with the NTK arises due to our choice of scaling in the ReLU layers. A different scaling leads to the mean field regime (Mei et al., 2018). The key insight of Mei et al., 2018 is that in the infinite-width limit, the gradient flow is captured by a specific non-linear partial differential equation (PDE). Due to non-asymptotic bounds on the accuracy of this measure-valued PDE model, new convergence results for (S)GD can be derived. The connection between the kernel and mean field regimes is explained in detail in Mei et al., 2019, Sec. 4 and App. H.

Random embeddings. A key component of our analysis is the capability of random ReLU-layers to preserve Euclidean distances with high probability (see Theorem 19). This finding is related to results on non-linear random embeddings, which play a major role in the field of quantized compressed sensing (e.g., see Jacques et al., 2013; Plan and Vershynin, 2014; Oymak and Recht, 2015; Cambareri et al., 2017; Dirksen and Mendelson, 2021; Dirksen, 2019; Xu and Jacques, 2020; Dirksen et al., 2022a,b). In particular, our choice of the bias vector (see Definition 1) is inspired by *dithering*, a technique that has already proven useful in various signal reconstruction problems (Jacques and Cambareri, 2017; Dirksen and Mendelson, 2018, 2021; Xu and Jacques, 2020; Jung et al., 2021). A remarkable new (and somewhat counterintuitive) insight of the present work is that for appropriate non-linearities like the ReLU-activation, desirable distance preservation properties and data separation can be achieved simultaneously.

Theorem 19 is new in its own right and improves on a previous result by Giryes et al. (2016), see also Giryes et al. (2020). It is also closely related to a work of Arpit and Bengio (2019), who have investigated the capability of a random ReLU-layer as in Definition 1 (but with bias $\boldsymbol{b} = \boldsymbol{0}$) to preserve Euclidean norms.

Rare eclipse problem. Finally, we point out an interesting connection between the separation capacity of random hyperplanes (see Theorem 6) and the rare eclipse problem studied by Bandeira et al. (2017); Cambareri et al. (2017). In both cases, the goal is to use a random transform $T : \mathbb{R}^d \to \mathbb{R}^k$ to map two linearly separable sets $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$ into a lower dimensional space $\mathbb{R}^k$ such that the following holds with a certain probability p:

$$T(\mathcal{X}^+) \cap T(\mathcal{X}^-) = \emptyset. \tag{14}$$

More specifically, the rare eclipse problem asks how small k can become such that (14) holds with probability at least $p = 1 - \eta$, where $\eta > 0$ is fixed but can be arbitrarily small. Using Gordon's Escape Through a Mesh Theorem (Gordon, 1988), Bandeira et al. (2017) have shown that if $\mathcal{X}^+$ and $\mathcal{X}^-$ are disjoint, closed, and convex sets, then $k \gtrsim w^2(\mathrm{cone}(\mathcal{X}^+ - \mathcal{X}^-) \cap \mathbb{S}^{d-1}) + \log(\eta^{-1})$ ensures (14) with probability at least $1 - \eta$, where $T \in \mathbb{R}^{k \times d}$ is a standard Gaussian random matrix.

In contrast, Theorem 6 considers a map of the form $T(\boldsymbol{x}) = \langle \boldsymbol{g}, \boldsymbol{x} \rangle + \tau$, where $\boldsymbol{g}$ is a standard Gaussian random vector and $\tau \in [-\lambda, \lambda]$ uniformly distributed for $\lambda > 0$ large enough. If $\mathcal{X}^+, \mathcal{X}^- \subset R\mathbb{B}_2^d$ are (ε, γ)-linearly separable (with some minimal distance), then (14) holds with probability at least $R\lambda^{-1} \exp(-Cw^2(\mathcal{X}^+ - \mathcal{X}^-))$, where $C > 0$ only depends on ε, γ, and R. Hence, Theorem 6 guarantees disjoint sets even for a single coordinate ($k = 1$), however at the expense of a worse probability of success. Remarkably, the Gaussian mean width and the difference set $\mathcal{X}^+ - \mathcal{X}^-$ play a key role both in the rare eclipse problem and Theorem 6.

1.6 Overview and Notation

The rest of the article is organized as follows: In Section 2, we present our main result, Theorem 10, based on the notion of mutual complexity (see Definition 8 and 9). The next two sections are then devoted to our main mathematical tools, namely separation by random hyperplanes (Section 3) and distance preservation (Section 4). Finally, the proof of Theorem 10 is given in Section 5, followed by a derivation of its variants (Theorem 2, 3, and 4) in Section 6.

Before proceeding, let us fix some standard notations and conventions that are commonly used in this paper. The letters c and C denote absolute (positive) constants, whose values may change from line to line. We speak of an *absolute constant* if its value does not depend on any other involved parameter. If an inequality holds up to an absolute constant C, we usually write $A \lesssim B$ instead of $A \leq C \cdot B$. The notation $A \asymp B$ is a shortcut for $A \lesssim B \lesssim A$.

For $d \in \mathbb{N}$, we set $[d] \coloneqq \{1, \dots, d\}$. The *cardinality* of an index set $I \subset [d]$ is denoted by $|I|$. Vectors and matrices are denoted by lower- and uppercase boldface letters, respectively. The i-th entry of a vector $\boldsymbol{z} \in \mathbb{R}^d$ is denoted by $[\boldsymbol{z}]_i$, or simply by z_i if there is no danger of confusion. We write $\boldsymbol{I}_d \in \mathbb{R}^{d \times d}$ and $\boldsymbol{0} \in \mathbb{R}^d$ for the *identity matrix* and the *zero vector* in $\mathbb{R}^d$, respectively. For $1 \leq q \leq \infty$, we denote the *ℓ^q-norm* on $\mathbb{R}^d$ by $\|\cdot\|_q$ and the associated closed *unit ball* by $\mathbb{B}_q^d$. The *Euclidean unit sphere* is given by $\mathbb{S}^{d-1} \coloneqq \{\boldsymbol{z} \in \mathbb{R}^d \mid \|\boldsymbol{z}\|_2 = 1\}$, and we also set $\mathbb{S}_+^{d-1} \coloneqq \mathbb{S}^{d-1} \cap [0, \infty)^d$.

Let $\mathcal{X}, \mathcal{X}' \subset \mathbb{R}^d$ and $\boldsymbol{z} \in \mathbb{R}^d$. The *linear cone* generated by $\mathcal{X}$ is denoted by $\operatorname{cone}(\mathcal{X}) \coloneqq \{v\tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{z}} \in \mathcal{X}, v \geq 0\}$. The *Minkowski difference* between $\mathcal{X}$ and $\mathcal{X}'$ is defined by $\mathcal{X} - \mathcal{X}' \coloneqq \{\boldsymbol{z}_1 - \boldsymbol{z}_2 \mid \boldsymbol{z}_1 \in \mathcal{X}, \boldsymbol{z}_2 \in \mathcal{X}'\}$, and we use the shortcut $\mathcal{X} - \boldsymbol{z} \coloneqq \mathcal{X} - \{\boldsymbol{z}\}$. The *distance* between $\boldsymbol{z}$ and $\mathcal{X}$ is $\operatorname{dist}(\boldsymbol{z}, \mathcal{X}) \coloneqq \inf_{\tilde{\boldsymbol{z}} \in \mathcal{X}} \|\boldsymbol{z} - \tilde{\boldsymbol{z}}\|_2$. Moreover, the *diameter* and *radius* of $\mathcal{X}$ are denoted by $\operatorname{diam}(\mathcal{X}) \coloneqq \sup_{\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{X}} \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2$ and $\operatorname{rad}(\mathcal{X}) \coloneqq \sup_{\tilde{\boldsymbol{z}} \in \mathcal{X}} \|\tilde{\boldsymbol{z}}\|_2$, respectively.

The *L^q-norm* of a real-valued random variable g is given by $\|g\|_{L^q} \coloneqq (\mathbb{E}[|g|^q])^{1/q}$. We call g *sub-Gaussian* if $\|g\|_{\psi_2} \coloneqq \inf\left\{v > 0 \mid \mathbb{E}[\exp(|g|^2/v^2)] \leq 2\right\} < \infty$; see Vershynin (2018, Chap. 2 & 3) for more details on sub-Gaussian random variables and their properties. Finally, we write $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ if $\boldsymbol{g}$ is a *standard Gaussian random vector* in $\mathbb{R}^d$.

The *ceiling* and *floor function* of $z \in \mathbb{R}$ are denoted by $\lceil z \rceil$ and $\lfloor z \rfloor$, respectively.

2. Main Separation Result and Mutual Complexity

This section presents our most general solution to Problem 1, containing all guarantees from the introduction (Theorem 2, 3, and 4) as special cases. To formulate the main result, Theorem 10, we require two important definitions formalizing the idea of *mutual complexity* between two sets. The first one can be seen as a refinement of the uniform covering introduced in (8):

Definition 8 (Mutual covering) *Let $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$ and $\lambda > 0$.*

*We call $\mathcal{C}^+ \coloneqq \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\} \subset \mathbb{R}^d$ and $\mathcal{C}^- \coloneqq \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\} \subset \mathbb{R}^d$ a λ-*mutual covering *for $\mathcal{X}^+$ and $\mathcal{X}^-$ if there exist $r_1^+, \dots, r_{N^+}^+ \geq 0$ and $r_1^-, \dots, r_{N^-}^- \geq 0$ such that*

(i) the sets $\mathcal{X}_j^+ \coloneqq \mathcal{X}^+ \cap \mathbb{B}_2^d(\boldsymbol{c}_j^+, r_j^+)$ for $j \in [N^+]$, and $\mathcal{X}_l^- \coloneqq \mathcal{X}^- \cap \mathbb{B}_2^d(\boldsymbol{c}_l^-, r_l^-)$ for $l \in [N^-]$, cover $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively;

(ii) $r_j^+ \leq \lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-)$ for all $j \in [N^+]$, and $r_l^- \leq \lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$ for all $l \in [N^-]$.

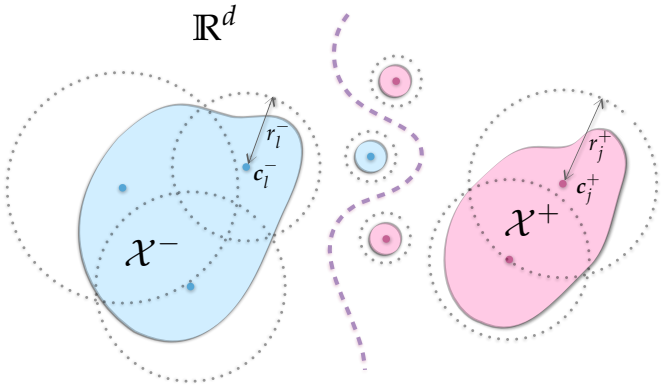

Figure 5: **Mutual covering.** This figure illustrates the geometric idea underlying Definition 8: those parts of $\mathcal{X}^-$ and $\mathcal{X}^+$ further away from the decision boundary may be covered by larger, and therefore fewer, Euclidean balls.

Furthermore, the sets $\mathcal{X}_1^+, \dots, \mathcal{X}_{N^+}^+ \subset \mathcal{X}^+$ *and* $\mathcal{X}_1^-, \dots, \mathcal{X}_{N^-}^- \subset \mathcal{X}^-$ *are referred to as the* components *of the covering.*

Although the notion of λ-mutual covering involves some technicalities, it is conceptually simple: We allow $\mathcal{X}^-$ and $\mathcal{X}^+$ to be covered by Euclidean balls of any radius, as long as the balls corresponding to different classes do not get too close in the sense of condition (ii). This constraint is also consistent with the setting of Theorem 4, which is obtained by choosing $r_j^+ = r_l^- = c\delta^2/\lambda$. However, Definition 8 is much more flexible and accounts for the mutual arrangement of the classes. For example, those parts of $\mathcal{X}^-$ that are far away from the decision boundary may be covered by a few large balls, while smaller radii are only needed for data closer to $\mathcal{X}^+$; see Figure 5 for an illustration. In general, this strategy leads to more efficient coverings and motivates the following geometric complexity parameters:

Definition 9 (Mutual complexity) *Let* $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$ *and* $\delta, \lambda > 0$.

We say that $\mathcal{X}^+$ *and* $\mathcal{X}^-$ *have* (R, δ, λ)-mutual complexity (N^+, N^-, w^+, w^-) *if there exists a* λ*-mutual covering* $\mathcal{C}^+ = \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\}$ *and* $\mathcal{C}^- = \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\}$ *for* $\mathcal{X}^+$ *and* $\mathcal{X}^-$ *such that*

(i) $\max_{j \in [N^+]} w(\mathcal{X}_j^+) \leq w^+$ *and* $\max_{l \in [N^-]} w(\mathcal{X}_l^-) \leq w^-$*;*

(ii) $\mathcal{C}^+, \mathcal{C}^- \subset R\mathbb{B}_2^d$ *are* δ*-separated.*

It is useful to keep in mind that the covering numbers N^+ and N^- reflect the *global* size of $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively, while w^+ and w^- should be viewed as *local* complexity measures (cf. (10)). In contrast, the parameters (R, δ, λ) are not instance-specific and concern the general problem setting.

We are now ready to state the main result of this work:

Theorem 10 (Main result) *There exist absolute constants* $c, C, C' > 0$ *such that the following holds.*

For $R \geq 1$, let $\mathcal{X}^-, \mathcal{X}^+ \subset R\mathbb{B}_2^d$ be δ-separated and let $\lambda \geq e\delta$ be such that $\lambda \gtrsim R\sqrt{\log(\lambda/\delta)}$. Furthermore, let $\mathcal{X}^+$ and $\mathcal{X}^-$ have $(R, \delta, C'\lambda)$-mutual complexity (N^+, N^-, w^+, w^-). We 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

$$\begin{aligned} n &\gtrsim w^2(\mathrm{cone}(\mathcal{X}^- - \mathcal{X}^-) \cap \mathbb{S}^{d-1}) + w^2(\mathrm{cone}(\mathcal{X}^+ - \mathcal{X}^+) \cap \mathbb{S}^{d-1}), \\ n &\gtrsim \left(\tfrac{\lambda}{\delta}\right)^8 \cdot \Big(\lambda^{-2}\big(w^- + w^+\big)^2 + \log(2N^-N^+/\eta)\Big) \end{aligned} \tag{15}$$

and

$$\begin{aligned} \hat{\lambda} &\gtrsim \left(\tfrac{\lambda}{\delta}\right)^4 \cdot \big(w^- + w(\mathcal{X}^+) + \lambda\big), \\ \hat{n} &\gtrsim \left(\tfrac{\hat{\lambda}}{w^- + w(\mathcal{X}^+) + \lambda}\right) \cdot \exp\Big(C \cdot \big(w^- + w(\mathcal{X}^+) + \lambda\big)^2 \cdot \lambda^6 \cdot \delta^{-8} \cdot \log(\lambda/\delta)\Big) \cdot \log(N^-/\eta). \end{aligned} \tag{16}$$

Then, given the two-layer random NN $F\colon \mathbb{R}^d \to \mathbb{R}^{\hat{n}}$, $\boldsymbol{x} \mapsto \hat{\Phi}(\Phi(\boldsymbol{x}))$, with probability at least $1 - \eta$, the sets $F(\mathcal{X}^-), F(\mathcal{X}^+) \subset \hat{\lambda}\mathbb{B}_2^{\hat{n}}$ are linearly separable with margin

$$c \cdot \tfrac{(w^- + w(\mathcal{X}^+) + \lambda)^2}{\hat{\lambda}} \cdot \exp\Big(-C \cdot \big(w^- + w(\mathcal{X}^+) + \lambda\big)^2 \cdot \lambda^6 \cdot \delta^{-8} \cdot \log(\lambda/\delta)\Big). \tag{17}$$

Despite a strong resemblance to Theorem 4, the above result entails several important improvements. First, the exponential terms in (16) and (17) only depend on the localized mean width w^-, but not the covering number N^-. Hence, the global size of $\mathcal{X}^-$ does not have any (negative) impact here. The situation is different for $\mathcal{X}^+$, whose complexity is still captured by $w(\mathcal{X}^+)$. In fact, the following adaption of (10) clarifies the role of N^+:

$$w(\mathcal{X}^+) \lesssim w^+ + R\sqrt{\log N^+}. \tag{18}$$

The aforementioned asymmetry in Theorem 10 becomes especially useful when the set $\mathcal{X}^+$ is relatively "small" compared to $\mathcal{X}^-$. A prototypical example in this regard is a low-complexity set ($= \mathcal{X}^+$), say a small Euclidean ball, which is surrounded by a hypersphere ($= \mathcal{X}^-$); see Figure 6 for an illustration.

Another distinctive feature of Theorem 10 is the usage of *mutual* complexity. To understand its merits over the uniform covering considered in Theorem 4, it is worth revisiting the scenario of Figure 5: while the largest portion of the two classes is away from the (δ-separated) decision boundary, only a few "outliers" are close to it. Thus, a uniform covering would preset a very small radius (at the order $O(\lambda^{-1}\delta^2)$), which is appropriate for the outlier part but inefficient for the remaining bulk; this would lead to unnecessarily large covering numbers and thereby to poor complexity bounds (cf. (18)). In contrast, our mutual covering strategy is flexible enough to handle such data configurations. Therefore, Theorem 10 indeed presents an *instance-specific* solution to Problem 1, including a variant of outlier robustness.

Remark 11 (Possible extensions) *For the sake of clarity, we have omitted some possible variations and generalizations of Theorem 10, which are however relatively straightforward to implement:*

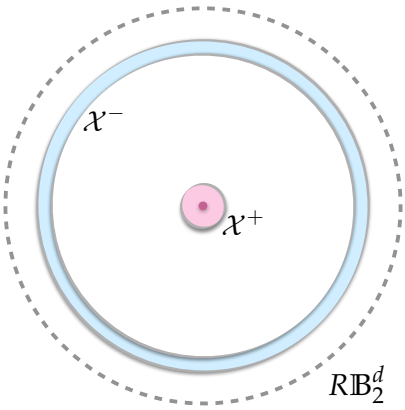

Figure 6: **An example of "asymmetric complexity" in the context of Theorem 10.** The set $\mathcal{X}^-$ corresponds to a thin hypersphere (say $\mathbb{S}^{d-1}$) around the origin, surrounding a small concentric ball $\mathcal{X}^+ = r\mathbb{B}_2^d$. If $r \lesssim 1/\sqrt{d}$, then w^- and $w(\mathcal{X}^+)$ are of constant order. Crucially, the covering number N^-, which scales exponentially in d, has no detrimental effect on the condition (16).

(1) Symmetry. *As discussed above, the asymmetric way of measuring complexity in Theorem 10 can be advantageous in certain situations. On the other hand, it is obvious that the roles of $\mathcal{X}^-$ and $\mathcal{X}^+$ are interchangeable. Hence, Theorem 10 could be "symmetrized" in this respect by a simple union bound argument.*

(2) Non-linear activation. *The considered random network design is tailored to the ReLU-activation (see Definition 1). Nevertheless, our proof strategy is applicable to other functions as well, e.g., the thresholding activation. This might involve a slight adaption of Definition 8(ii) and lead to a different scaling of δ and λ in Theorem 10, but the qualitative statement remains valid.*

(3) Multiclass classification. *While we have focused on binary labels for the sake of simplicity, our main results can readily be extended to categorical data using a simple one-vs-rest strategy. Assume we are given data from K different classes, say $\mathcal{X}^1, \mathcal{X}^2, \dots, \mathcal{X}^K \subset \mathbb{R}^d$. Then, for any $l \in [K]$, Theorem 10 implies that a sufficiently large random NN separates $\mathcal{X}^+ := \mathcal{X}^l$ and $\mathcal{X}^- := \bigcup_{k\in[K]\setminus\{l\}} \mathcal{X}^k$ with high probability. Taking the union bound over these K events, we conclude that with high probability a single, large random NN F makes each individual set $F(\mathcal{X}^1), \dots, F(\mathcal{X}^K)$ linearly separable from the remaining ones. Analogously to the binary case, this separation property allows us to train a standard one-vs-rest SVM classifier on the transformed data sets.*

3. Separation by Random Hyperplanes

The goal of this section is to prove Theorem 6, which is our main result on the separation of two sets by a random hyperplane. Before outlining the main steps of our proof, let us define the relevant notions of separability.

Definition 12 *Let $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$.*

(a) Let $\boldsymbol{v} \in \mathbb{R}^d \backslash \{\boldsymbol{0}\}, \tau \in \mathbb{R}$ and $t \geq 0$. A hyperplane $H[\boldsymbol{v}, \tau]$ t-separates $\mathcal{X}^-$ from $\mathcal{X}^+$ *if*

$$\begin{aligned} \langle \boldsymbol{v}, \boldsymbol{x}^- \rangle + \tau &\leq -t \qquad \text{for all } \boldsymbol{x}^- \in \mathcal{X}^-, \\ \langle \boldsymbol{v}, \boldsymbol{x}^+ \rangle + \tau &> +t \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+. \end{aligned}$$

If $t = 0$, we simply say that $H[\boldsymbol{v}, \tau]$ separates $\mathcal{X}^-$ from $\mathcal{X}^+$.

(b) Let $\varepsilon \in [0, 1]$ and $\gamma > 0$. We say that $\mathcal{X}^+$ and $\mathcal{X}^-$ are (ε, γ)-linearly separable *if $\mathcal{X}^+$ and $\mathcal{X}^-$ are γ-separated (see Problem 1) and there exists $\boldsymbol{u} \in \mathbb{S}^{d-1}$ such that*

$$\langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq (1 - \varepsilon) \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+ \text{ and } \boldsymbol{x}^- \in \mathcal{X}^-.$$

Recall from (1) that $\mathcal{X}^+$ and $\mathcal{X}^-$ are called *linearly separable with margin t* if they are t-separated by a hyperplane $H[\boldsymbol{v}, \tau]$ with $\|\boldsymbol{v}\|_2 = 1$. In comparison, (ε, γ)-linearly separability is a strictly stronger condition (see also Proposition 13(iii) below). Intuitively, it captures how much a separating hyperplane can be perturbed, such that it still separates the sets $\mathcal{X}^+$ and $\mathcal{X}^-$; geometrically, the parameter ε controls the narrowness of $\operatorname{cone}(\mathcal{X}^+ - \mathcal{X}^-)$.

Proof sketch for Theorem 6. By a rescaling argument, we can assume that $R = 1$. For a $k \in \mathbb{N}$ specified below, we represent the standard Gaussian vector $\boldsymbol{g} \in \mathbb{R}^d$ by $\boldsymbol{g} = \boldsymbol{G}^T \boldsymbol{v}'$, where $\boldsymbol{G} \in \mathbb{R}^{k \times d}$ is a standard Gaussian matrix, $\boldsymbol{v}' \in \mathbb{S}^{k-1}$ is uniformly distributed, and $\boldsymbol{G}, \boldsymbol{v}'$ are independent. We then observe that, for any $\rho \geq 0$, the hyperplane $H[\boldsymbol{g}, \tau]$ ρ-separates $\mathcal{X}^-$ from $\mathcal{X}^+$ if and only if the hyperplane $H[\sqrt{k}\boldsymbol{v}', \tau]$ ρ-separates $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^-$ from $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^+$. Therefore, one can prove Theorem 6 by first showing that for k large enough, the sets $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^-$ and $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^+$ are again linearly separable with constant probability and second, showing that conditioned on this event the hyperplane $H[\sqrt{k}\boldsymbol{v}', \tau]$ ρ-separates $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^-$ and $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^+$ with probability p, where ρ and p are specified in Theorem 6. Specifically, the main technical steps are:

1. to show that if
$$k \gtrsim \gamma^{-2}(1 - \varepsilon)^{-2}(w^2(\mathcal{X}^+ - \mathcal{X}^-) + 1),$$
then the linear transformation $\frac{1}{\sqrt{k}}\boldsymbol{G}$ maps the (ε, γ)-linearly separable sets $\mathcal{X}^-$ and $\mathcal{X}^+$ to $(\frac{1+\varepsilon}{2}, \frac{\gamma}{2})$-linearly separable sets $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^-$ and $\frac{1}{\sqrt{k}}\boldsymbol{G}\mathcal{X}^+$ with probability at least $\frac{1}{2}$.
2. to derive a general separation result for two (ε, γ)-linearly separable sets by a random hyperplane $H[\boldsymbol{v}, \tau]$, where $\boldsymbol{v}$ is uniformly distributed on Euclidean sphere (see Theorem 14 and Corollary 15).

Let us now give the proof in full detail. We start with a simple proposition that relates our notions of separability.

Proposition 13 *Let $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$. The following relationships hold:*

(i) If $\mathcal{X}^+$ and $\mathcal{X}^-$ are linearly separable with margin μ, then they are 2μ-separated.

(ii) *If a hyperplane $H[\boldsymbol{u},\tau]$ t-separates $\mathcal{X}^-$ from $\mathcal{X}^+$, then $\mathcal{X}^+$ and $\mathcal{X}^-$ are linearly separable with margin $t/\|\boldsymbol{u}\|_2$.*

(iii) *If $\mathcal{X}^+$ and $\mathcal{X}^-$ are (ε,γ)-linearly separable, then they are linearly separable with margin $\frac{(1-\varepsilon)\gamma}{2}$.*

(iv) *If $\mathcal{X}^+$ and $\mathcal{X}^-$ are linearly separable with margin μ and $\operatorname{diam}(\mathcal{X}^+ - \mathcal{X}^-) \leq R$, then they are $(\frac{R-2\mu}{R}, 2\mu)$-linearly separable.*

Proof To show (i), observe that by assumption there exist $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and $\tau \in \mathbb{R}$ such that

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

It follows that $\langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq 2\mu$ for all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$. By the Cauchy-Schwarz inequality, $\langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \leq \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$, which shows the claim. For (ii), it suffices to note that if $H[\boldsymbol{u},\tau]$ t-separates $\mathcal{X}^-$ from $\mathcal{X}^+$, then $H[\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}, \frac{\tau}{\|\boldsymbol{u}\|_2}]$ $\frac{t}{\|\boldsymbol{u}\|_2}$-separates $\mathcal{X}^-$ from $\mathcal{X}^+$. Let us next show (iii). If $\mathcal{X}^+$ and $\mathcal{X}^-$ are (ε,γ)-linearly separable, then there exists $\boldsymbol{u} \in \mathbb{S}^{d-1}$ such that

$$\langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq (1-\varepsilon)\gamma \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+ \text{ and } \boldsymbol{x}^- \in \mathcal{X}^-.$$

Set $\tau = -\frac{(1-\varepsilon)\gamma}{2} + \inf_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{u}, -\boldsymbol{x}^- \rangle$. Fix $\boldsymbol{x}' \in \mathcal{X}^+$. Since

$$\inf_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{u}, -\boldsymbol{x}^- \rangle = -\langle \boldsymbol{u}, \boldsymbol{x}' \rangle + \inf_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{u}, \boldsymbol{x}' - \boldsymbol{x}^- \rangle \geq -\langle \boldsymbol{u}, \boldsymbol{x}' \rangle + (1-\varepsilon)\gamma,$$

we see that $\tau \in \mathbb{R}$. Further, for any $\boldsymbol{x}^- \in \mathcal{X}^-$,

$$\begin{aligned} \langle \boldsymbol{u}, \boldsymbol{x}^- \rangle + \tau &= \langle \boldsymbol{u}, \boldsymbol{x}^- \rangle - \tfrac{(1-\varepsilon)\gamma}{2} + \inf_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{u}, -\boldsymbol{x}^- \rangle \\ &= \langle \boldsymbol{u}, \boldsymbol{x}^- \rangle - \sup_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{u}, \boldsymbol{x}^- \rangle - \tfrac{(1-\varepsilon)\gamma}{2} \leq -\tfrac{(1-\varepsilon)\gamma}{2} \end{aligned}$$

and for any $\boldsymbol{x}^+ \in \mathcal{X}^+$,

$$\begin{aligned} \langle \boldsymbol{u}, \boldsymbol{x}^+ \rangle + \tau &= \langle \boldsymbol{u}, \boldsymbol{x}^+ \rangle - \tfrac{(1-\varepsilon)\gamma}{2} + \inf_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{u}, -\boldsymbol{x}^- \rangle \\ &= \inf_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle - \tfrac{(1-\varepsilon)\gamma}{2} \geq \tfrac{(1-\varepsilon)\gamma}{2}. \end{aligned}$$

Since $\|\boldsymbol{u}\|_2 = 1$, it follows that the hyperplane $H[\boldsymbol{u},\tau]$ linearly separates $\mathcal{X}^-$ and $\mathcal{X}^+$ with margin $\frac{(1-\varepsilon)\gamma}{2}$. Finally, let us show (iv). By (i) we know that $\mathcal{X}^+$ and $\mathcal{X}^-$ are 2μ-separated. Let $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and $\tau \in \mathbb{R}$ be such that

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

It follows that $\langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq 2\mu$ for all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$. Since $\operatorname{diam}(\mathcal{X}^+ - \mathcal{X}^-) \leq R$, we also have $2\mu \geq \frac{2\mu}{R}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$ for all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$. Together this yields

$$\langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq \big(1 - \tfrac{R-2\mu}{R}\big)\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+ \text{ and } \boldsymbol{x}^- \in \mathcal{X}^-. \qquad \blacksquare$$

The next result gives a lower bound for the probability that a random hyperplane $H[\boldsymbol{v}, \tau]$ separates two (ε, γ)-linearly separable sets $\mathcal{X}^+, \mathcal{X}^- \subset R\mathbb{B}_2^d$, where $\tau \in [-\lambda, \lambda]$ is uniformly distributed for $\lambda \geq R$ and $\boldsymbol{v}$ is uniformly distributed on $\mathbb{S}^{d-1}$. As detailed in our above proof sketch, this result (more precisely, Corollary 15) forms a crucial ingredient of our proof of Theorem 6.

Theorem 14 *There exist absolute constants $c, C > 0$ such that the following holds.*

For $\varepsilon \in [0,1]$ and $\gamma > 0$, consider (ε, γ)-linearly separable sets $\mathcal{X}^+, \mathcal{X}^- \subset R\mathbb{B}_2^d$. Let $\boldsymbol{v} \in \mathbb{S}^{d-1}$ and $\tau \in [-\lambda, \lambda]$ be both uniformly distributed. If $\lambda \geq R$, then with probability at least

$$c\tfrac{\gamma}{\lambda}(1-\varepsilon)(\sqrt{\varepsilon} + \tfrac{1}{\sqrt{d}}) \exp(-C\varepsilon d \log(2(1-\varepsilon)^{-1})),$$

the hyperplane $H[\boldsymbol{v}, \tau]$ t-separates $\mathcal{X}^-$ from $\mathcal{X}^+$ with $t = c\gamma(1-\varepsilon)(\sqrt{\varepsilon} + \frac{1}{\sqrt{d}})$.

Corollary 15 *There exist absolute constants $c, C > 0$ such that the following holds.*

For $\varepsilon \in [0,1]$ and $\gamma > 0$, consider (ε, γ)-linearly separable sets $\mathcal{X}^+, \mathcal{X}^- \subset R\mathbb{B}_2^d$. Let $\nu > 0$. Let $\boldsymbol{v} \in \mathbb{S}^{d-1}$ and $\tau \in [-\lambda, \lambda]$ be both uniformly distributed. If $\lambda \geq \nu R$, then with probability at least

$$c\tfrac{\nu\gamma}{\lambda}(1-\varepsilon)(\sqrt{\varepsilon} + \tfrac{1}{\sqrt{d}}) \exp(-C\varepsilon d \log(2(1-\varepsilon)^{-1})),$$

the hyperplane $H[\nu\boldsymbol{v}, \tau]$ t-separates $\mathcal{X}^-$ from $\mathcal{X}^+$ with $t = c\gamma(1-\varepsilon)(\sqrt{\varepsilon} + \frac{1}{\sqrt{d}})\nu$.

Proof For $t \geq 0$ the hyperplane $H[\nu\boldsymbol{v}, \tau]$ $t\nu$-separates $\mathcal{X}^-$ from $\mathcal{X}^+$ if and only if the hyperplane $H[\boldsymbol{v}, \frac{1}{\nu}\tau]$ t-separates $\mathcal{X}^-$ from $\mathcal{X}^+$. The random variable $\tau' := \frac{1}{\nu}\tau$ is uniformly distributed on $[-\lambda', \lambda']$ for $\lambda' = \frac{\lambda}{\nu}$. The result follows from Theorem 14 for $t = c\gamma(1-\varepsilon)(\sqrt{\varepsilon} + \frac{1}{\sqrt{d}})$. ■

For the proof of Theorem 14, we need the following standard result (e.g., see Boucheron et al., 2013, Sec. 7.2), which precisely describes the surface measure of a spherical cap.

Lemma 16 *Let $\boldsymbol{v} \in \mathbb{S}^{d-1}$ be uniformly distributed. Let $\delta \in (0,1]$ and $d \geq 2\delta^{-2}$. For any $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we have that*

$$\tfrac{1}{6\delta\sqrt{d}}(1-\delta^2)^{\frac{d-1}{2}} \leq \mathbb{P}\left(\langle \boldsymbol{v}, \boldsymbol{u}\rangle \geq \delta\right) \leq \tfrac{1}{2\delta\sqrt{d}}(1-\delta^2)^{\frac{d-1}{2}}.$$

If additionally $\delta \leq \frac{1}{\sqrt{2}}$, then

$$\mathbb{P}\left(\langle \boldsymbol{v}, \boldsymbol{u}\rangle \geq \delta\right) \geq \tfrac{1}{2}\exp(-2\delta^2 d).$$

Proof [Theorem 14] For $\theta \geq 0$ define the event[8]

$$\mathsf{A}_{\boldsymbol{v}}(\theta) := \Big\{ \inf_{\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{v}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq \theta \Big\}.$$

8. Formally, all events should be understood in the ordinary sense of probability theory, i.e., measurable subsets of some appropriate sample space. Note that the underlying probability space is not explicitly mentioned here. Our analysis does not require any treatment of measure theoretic issues, and we simply assume that the probability space is rich enough to model all random quantities and processes that we are interested in.

For any $s \geq 0$, we have that

$$\begin{aligned}&\mathbb{P}(H[\boldsymbol{v},\tau] \text{ } s\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+)\\ &\geq \mathbb{P}(\{H[\boldsymbol{v},\tau] \text{ } s\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+\} \cap \mathsf{A}_{\boldsymbol{v}}(\theta))\\ &= p_1(s,\theta)\cdot p_2(\theta),\end{aligned}$$

where

$$p_1(s,\theta) \coloneqq \mathbb{P}(H[\boldsymbol{v},\tau] \text{ } s\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+ \mid \mathsf{A}_{\boldsymbol{v}}(\theta)) \quad \text{and} \quad p_2(\theta) \coloneqq \mathbb{P}(\mathsf{A}_{\boldsymbol{v}}(\theta)).$$

Next, we bound both factors $p_1(s,\theta)$ and $p_2(\theta)$ from below.

Lower bound for $p_1(s,\theta)$. Let us show that if $\lambda \geq R$, then for $s \leq \frac{\theta}{2}$,

$$\mathbb{P}(H[\boldsymbol{v},\tau] \text{ } s\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+ \mid \mathsf{A}_{\boldsymbol{v}}(\theta)) \geq \tfrac{\theta-2s}{2\lambda}.$$

If $\lambda \geq R$, then $a \coloneqq \sup_{\boldsymbol{x}^- \in \mathcal{X}^-} \langle \boldsymbol{v}, \boldsymbol{x}^- \rangle \in [-\lambda, \lambda]$ and $b \coloneqq \inf_{\boldsymbol{x}^+ \in \mathcal{X}^+} \langle \boldsymbol{v}, \boldsymbol{x}^+ \rangle \in [-\lambda, \lambda]$. Further, on the event $\mathsf{A}_{\boldsymbol{v}}(\theta)$ it holds $b - a \geq \theta$. Let $s \leq \frac{\theta}{2}$. If $-\tau \in [a+s, b-s]$, then

$$\begin{aligned}\langle \boldsymbol{v}, \boldsymbol{x}^- \rangle + \tau \leq -s \qquad &\text{for all } \boldsymbol{x}^- \in \mathcal{X}^-,\\ \langle \boldsymbol{v}, \boldsymbol{x}^+ \rangle + \tau \geq +s \qquad &\text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+.\end{aligned}$$

Since $\mathbb{P}_\tau(-\tau \in [a+s, b-s]) \geq \frac{\theta-2s}{2\lambda}$, this shows

$$\mathbb{P}_\tau(H[\boldsymbol{v},\tau] \text{ } s\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+ \mid \mathsf{A}_{\boldsymbol{v}}(\theta)) \geq \tfrac{\theta-2s}{2\lambda}.$$

For $s = \frac{\theta}{4}$ we obtain

$$\mathbb{P}(H[\boldsymbol{v},\tau] \text{ } \tfrac{\theta}{4}\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+) \geq \tfrac{\theta}{4\lambda} \cdot p_2(\theta). \tag{19}$$

Lower bound for $p_2(\theta)$. Set $\mathcal{X} \coloneqq \operatorname{cone}(\mathcal{X}^+ - \mathcal{X}^-) \cap \mathbb{S}^{d-1}$ and for $\delta \in [0,1]$ define

$$\alpha_\varepsilon(\delta) \coloneqq \max\{\delta - \sqrt{2\varepsilon}, 1 - \varepsilon - \sqrt{2}\sqrt{1-\delta}\}.$$

Since $\mathcal{X}^+$ and $\mathcal{X}^-$ are (ε,γ)-linearly separable there exists $\boldsymbol{u} \in \mathbb{S}^{d-1}$ such that

$$\langle \boldsymbol{u}, \boldsymbol{z} \rangle \geq 1 - \varepsilon \qquad \text{for all } \boldsymbol{z} \in \mathcal{X}. \tag{20}$$

Let us show that for any $\delta \in [0,1]$ with $\alpha_\varepsilon(\delta) \geq 0$,

$$\{\langle \boldsymbol{v}, \boldsymbol{u} \rangle \geq \delta\} \subset \mathsf{A}_{\boldsymbol{v}}(\alpha_\varepsilon(\delta)\gamma). \tag{21}$$

First observe that for any $\delta \in [0,1]$,

$$\{\langle \boldsymbol{v}, \boldsymbol{u} \rangle \geq \delta\} \subset \Big\{ \inf_{\boldsymbol{z} \in \mathcal{X}} \langle \boldsymbol{v}, \boldsymbol{z} \rangle \geq \alpha_\varepsilon(\delta) \Big\}. \tag{22}$$

Indeed, by (20), $\|\boldsymbol{z} - \boldsymbol{u}\|_2 \leq \sqrt{2\varepsilon}$ for every $\boldsymbol{z} \in \mathcal{X}$. Since $\boldsymbol{v} \in \mathbb{S}^{d-1}$ it follows

$$\langle \boldsymbol{v}, \boldsymbol{z} \rangle \geq \langle \boldsymbol{v}, \boldsymbol{u} \rangle - \|\boldsymbol{z} - \boldsymbol{u}\|_2 \geq \langle \boldsymbol{v}, \boldsymbol{u} \rangle - \sqrt{2\varepsilon} \geq \delta - \sqrt{2\varepsilon} \quad \text{for all } \boldsymbol{z} \in \mathcal{X}, \tag{23}$$

if $\langle \boldsymbol{v}, \boldsymbol{u} \rangle \geq \delta$. Moreover, $\langle \boldsymbol{v}, \boldsymbol{u} \rangle \geq \delta \Leftrightarrow \|\boldsymbol{v} - \boldsymbol{u}\|_2 \leq \sqrt{2}\sqrt{1-\delta}$. Therefore, if $\langle \boldsymbol{v}, \boldsymbol{u} \rangle \geq \delta$ then for every $\boldsymbol{z} \in \mathcal{X}$,

$$\langle \boldsymbol{v}, \boldsymbol{z} \rangle \geq \langle \boldsymbol{u}, \boldsymbol{z} \rangle - \|\boldsymbol{v} - \boldsymbol{u}\|_2 \geq 1 - \varepsilon - \sqrt{2}\sqrt{1-\delta}. \tag{24}$$

Inequalities (23) and (24) imply (22). If $\inf_{\boldsymbol{z} \in \mathcal{X}} \langle \boldsymbol{v}, \boldsymbol{z} \rangle \geq \alpha_\varepsilon(\delta)$ and $\alpha_\varepsilon(\delta) \geq 0$, the following holds for every $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$:

$$\langle \boldsymbol{v}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq \alpha_\varepsilon(\delta) \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \geq \alpha_\varepsilon(\delta)\gamma,$$

where for the second inequality we used that $\alpha_\varepsilon(\delta) \geq 0$ and that $\mathcal{X}^+$ and $\mathcal{X}^-$ are γ-separated. In combination with (22) this shows (21). From (21) it follows that for every $\delta \in [0,1]$ with $\alpha_\varepsilon(\delta) \geq 0$ and every $\theta \leq \alpha_\varepsilon(\delta)\gamma$,

$$p_2(\theta) \geq \mathbb{P}(\langle \boldsymbol{v}, \boldsymbol{u} \rangle \geq \delta). \tag{25}$$

In order to bound the probability on the right hand side from below, we distinguish the cases $\varepsilon \leq \frac{1}{32}$ and $\varepsilon > \frac{1}{32}$.

Case $\boldsymbol{\varepsilon \leq \frac{1}{32}}$. If $\delta \in [0,1]$ satisfies $\delta \geq \sqrt{8\varepsilon}$, then $\delta - \sqrt{2\varepsilon} \geq \frac{\delta}{2}$, which implies $\alpha_\varepsilon(\delta) \geq \frac{\delta}{2} \geq 0$. Therefore, by (25) the following holds: For all $\delta \in [\sqrt{8\varepsilon}, 1]$,

$$p_2(\tfrac{\delta}{2}\gamma) \geq \mathbb{P}(\langle \boldsymbol{v}, \boldsymbol{u} \rangle \geq \delta).$$

By Lemma 16 we obtain that for all $\delta \in (\sqrt{8\varepsilon}, \frac{1}{\sqrt{2}}]$ with $d \geq 2\delta^{-2}$,

$$p_2(\tfrac{\delta}{2}\gamma) \geq \tfrac{1}{2}\exp(-2\delta^2 d).$$

Applying (19) for $\theta = \frac{\delta}{2}\gamma$, we obtain that for all $\delta \in (\sqrt{8\varepsilon}, \frac{1}{\sqrt{2}}]$ with $d \geq 2\delta^{-2}$,

$$\mathbb{P}(H[\boldsymbol{v}, \tau] \ \tfrac{\delta\gamma}{8}\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+) \geq \tfrac{\delta\gamma}{8\lambda} \cdot \tfrac{1}{2}\exp(-2\delta^2 d).$$

The choice $\delta = \sqrt{8\varepsilon} + \sqrt{\frac{2}{d}}$, which satisfies $\delta \in (\sqrt{8\varepsilon}, \frac{1}{\sqrt{2}}]$ and $d \geq 2\delta^{-2}$, yields

$$\mathbb{P}(H[\boldsymbol{v}, \tau] \ \tfrac{\gamma}{8}(\sqrt{8\varepsilon} + \sqrt{\tfrac{2}{d}})\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+) \geq c\tfrac{\gamma}{\lambda}(\sqrt{\varepsilon} + \tfrac{1}{\sqrt{d}}) \cdot \exp(-C\varepsilon d)$$

for absolute constants $c, C > 0$. The result in the case $\varepsilon \leq \frac{1}{32}$ follows by observing that $1 - \varepsilon \sim 1$.

Case $\boldsymbol{\varepsilon > \frac{1}{32}}$. Set $\delta' = 1 - \frac{1}{4}(1-\varepsilon)^2$. Then $\delta' \in [0,1]$ and $\alpha_\varepsilon(\delta') \geq 1 - \varepsilon - \sqrt{2}\sqrt{1-\delta'} = (1 - \frac{1}{\sqrt{2}})(1-\varepsilon) \geq \frac{1}{4}(1-\varepsilon) \geq 0$. Therefore, by (25) in combination with Lemma 16 we obtain

$$p_2(\tfrac{1}{4}(1-\varepsilon)\gamma) \geq \tfrac{1}{6\delta'\sqrt{d}}(1-\delta'^2)^{\frac{d-1}{2}}.$$

Observe that

$$\begin{aligned}\tfrac{1}{6\delta'\sqrt{d}}(1-\delta'^2)^{\frac{d-1}{2}} &\geq \tfrac{1}{6\sqrt{d}}\exp(-d\log((1-\delta'^2)^{-1})) \\ &\geq \tfrac{1}{6\sqrt{d}}\exp(-d\log((1-\delta')^{-1}))\end{aligned}$$

$$= \tfrac{1}{6\sqrt{d}} \exp(-2d \log(2(1-\varepsilon)^{-1})) \\ \geq \exp(-3d \log(2(1-\varepsilon)^{-1})).$$

Applying (19) for $\theta = \frac{1}{4}(1-\varepsilon)\gamma$, we obtain

$$\mathbb{P}(H[\boldsymbol{v}, \tau] \ \tfrac{(1-\varepsilon)\gamma}{16}\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+) \geq \tfrac{(1-\varepsilon)\gamma}{16\lambda} \cdot \exp(-3d \log(2(1-\varepsilon)^{-1})).$$

The result in the case $\varepsilon > \frac{1}{32}$ follows by observing that $\sqrt{\varepsilon} + \frac{1}{\sqrt{d}} \sim 1$ and $\varepsilon \sim 1$. This completes the proof. ■

3.1 Proof of Theorem 6

For the proof of Theorem 6, we need one final ingredient, namely Lemma 17 below. It gives a sufficient condition under which a set contained in a spherical cone is again contained in a spherical cone after a linear transformation. Using this lemma, we will show that a Gaussian matrix with enough rows maps (ε, γ)-linear separable sets to $(\frac{1+\varepsilon}{2}, \frac{\gamma}{2})$-linearly separable sets with constant probability (see step 1 of the proof sketch at the beginning of Section 3).

Lemma 17 *Let $\boldsymbol{u} \in \mathbb{S}^{d-1}$, $t \in [0,1]$, and $\mathcal{X} \subset R\mathbb{B}_2^d$ satisfy*

$$\langle \boldsymbol{u}, \boldsymbol{x} \rangle \geq t\|\boldsymbol{x}\|_2 \qquad \text{for all } \boldsymbol{x} \in \mathcal{X}.$$

For $\kappa \in (0, \frac{1}{2}]$, $\alpha, \beta \in [0, R]$ and $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ assume that the following holds:

(i) $(1-\kappa)\|\boldsymbol{u}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{u}\|_2^2 \leq (1+\kappa)\|\boldsymbol{u}\|_2^2$,

(ii) $\|\boldsymbol{u} - \boldsymbol{x}\|_2 - \alpha \leq \|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{u} - \boldsymbol{x}\|_2 + \alpha$ *for all* $\boldsymbol{x} \in \mathcal{X}$,

(iii) $\|\boldsymbol{x}\|_2 - \beta \leq \|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_2 + \beta$ *for all* $\boldsymbol{x} \in \mathcal{X}$.

Then, we have that $\boldsymbol{A}\mathcal{X} \subset 2R\mathbb{B}_2^k$ and

$$\langle \tfrac{\boldsymbol{A}\boldsymbol{u}}{\|\boldsymbol{A}\boldsymbol{u}\|_2}, \boldsymbol{A}\boldsymbol{x} \rangle \geq \tfrac{t}{\sqrt{1+\kappa}}\|\boldsymbol{A}\boldsymbol{x}\|_2 - \tfrac{\kappa}{\sqrt{2}} - \sqrt{2}(\tfrac{3}{2}R + t)\beta - \tfrac{3}{\sqrt{2}}(1+R)\alpha \qquad \text{for all } \boldsymbol{x} \in \mathcal{X}. \quad (26)$$

Proof For any $\boldsymbol{x} \in \mathcal{X}$,

$$\big|\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\big| = \big|\|\boldsymbol{A}\boldsymbol{x}\|_2 - \|\boldsymbol{x}\|_2\big| \cdot \big|\|\boldsymbol{A}\boldsymbol{x}\|_2 + \|\boldsymbol{x}\|_2\big| \leq \beta(2\|\boldsymbol{x}\|_2 + \beta) \leq 3R\beta.$$

Analogously,

$$\big|\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{A}\boldsymbol{x}\|_2^2 - \|\boldsymbol{u} - \boldsymbol{x}\|_2^2\big| \leq 3(1+R)\alpha.$$

Therefore,

$$\begin{aligned} \langle \boldsymbol{A}\boldsymbol{u}, \boldsymbol{A}\boldsymbol{x} \rangle &= \tfrac{1}{2}\|\boldsymbol{A}\boldsymbol{u}\|_2^2 + \tfrac{1}{2}\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - \tfrac{1}{2}\|\boldsymbol{A}(\boldsymbol{u} - \boldsymbol{x})\|_2^2 \\ &\geq \tfrac{1}{2}(1-\kappa)\|\boldsymbol{u}\|_2^2 + \tfrac{1}{2}\|\boldsymbol{x}\|_2^2 - \tfrac{3}{2}R\beta - \tfrac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2 - \tfrac{3}{2}(1+R)\alpha \\ &= \langle \boldsymbol{u}, \boldsymbol{x} \rangle - \tfrac{\kappa}{2}\|\boldsymbol{u}\|_2^2 - \tfrac{3}{2}R\beta - \tfrac{3}{2}(1+R)\alpha \end{aligned}$$

$$\geq t\|\boldsymbol{x}\|_2 - \tfrac{\kappa}{2} - \tfrac{3}{2}R\beta - \tfrac{3}{2}(1+R)\alpha$$
$$\geq t\|\boldsymbol{A}\boldsymbol{x}\|_2 - \tfrac{\kappa}{2} - (\tfrac{3}{2}R + t)\beta - \tfrac{3}{2}(1+R)\alpha.$$

Using $\frac{1}{\sqrt{2}} \leq \|\boldsymbol{A}\boldsymbol{u}\|_2 \leq \sqrt{1+\kappa}$, we obtain (26). ■

We are now ready to prove our main result on the separation of two sets by a random hyperplane:

Proof [Theorem 6] First note that it suffices to prove the result for $R = 1$. Indeed, the general result then follows by a rescaling argument.

Let $k \in \mathbb{N}$. Let $\boldsymbol{v}' \in \mathbb{S}^{k-1}$ be uniformly distributed, $\boldsymbol{G} \in \mathbb{R}^{k\times d}$ a standard Gaussian matrix and $\tau \in [-\lambda, \lambda]$ be uniformly distributed. Let all random variables be independent. We define the random vector $\boldsymbol{g} := \boldsymbol{G}^T\boldsymbol{v}' \in \mathbb{R}^d$ and observe that it is standard Gaussian. Indeed, one may write $\boldsymbol{v}' = \boldsymbol{Q}\boldsymbol{e}_1$ where $\boldsymbol{Q} \in \mathbb{R}^{k\times k}$ is a uniform random orthogonal matrix and $\boldsymbol{e}_1$ the first unit vector in $\mathbb{R}^k$. Due to the rotational invariance of standard Gaussian matrices, we have that $\boldsymbol{G}^T\boldsymbol{Q} \sim \boldsymbol{G}^T$, and therefore, $\boldsymbol{g} = \boldsymbol{G}^T\boldsymbol{v}' = \boldsymbol{G}^T\boldsymbol{Q}\boldsymbol{e}_1 \sim \boldsymbol{G}^T\boldsymbol{e}_1 \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$.

Set $\boldsymbol{A} = \frac{1}{\sqrt{k}}\boldsymbol{G}$. For $\varepsilon' \in [0,1]$, and $\gamma' \geq 0$, we define the event

$$\mathsf{A}_{\varepsilon',\gamma'} := \{\boldsymbol{A}\mathcal{X}^+, \boldsymbol{A}\mathcal{X}^- \subset 5\mathbb{B}_2^k \text{ are } (\varepsilon',\gamma')\text{-linearly separable}\}.$$

Using that $\boldsymbol{g} = \boldsymbol{A}^T\sqrt{k}\boldsymbol{v}'$, we obtain for any $\rho \geq 0$,

$$\begin{aligned}
&\mathbb{P}(H[\boldsymbol{g},\tau] \text{ } \rho\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+) \\
&= \mathbb{P}(H[\sqrt{k}\boldsymbol{v}',\tau] \text{ } \rho\text{-separates } \boldsymbol{A}\mathcal{X}^- \text{ from } \boldsymbol{A}\mathcal{X}^+) \\
&\geq \mathbb{P}(\{H[\sqrt{k}\boldsymbol{v}',\tau] \text{ } \rho\text{-separates } \boldsymbol{A}\mathcal{X}^- \text{ from } \boldsymbol{A}\mathcal{X}^+\} \cap \mathsf{A}_{\varepsilon',\gamma'}) \\
&= \mathbb{P}(H[\sqrt{k}\boldsymbol{v}',\tau] \text{ } \rho\text{-separates } \boldsymbol{A}\mathcal{X}^- \text{ from } \boldsymbol{A}\mathcal{X}^+ \mid \mathsf{A}_{\varepsilon',\gamma'}) \cdot \mathbb{P}(\mathsf{A}_{\varepsilon',\gamma'}).
\end{aligned} \tag{27}$$

Since $\mathcal{X}^+$ and $\mathcal{X}^-$ are (ε,γ)-linearly separable, there exists a vector $\boldsymbol{u} \in \mathbb{S}^{d-1}$ such that

$$\langle \boldsymbol{u}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq (1-\varepsilon)\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-.$$

For $\kappa \in (0, \frac{1}{2}]$, $\alpha \in [0,2]$, $\beta \in [0, \frac{\gamma}{2}]$ define the event $\mathsf{B}_{\kappa,\alpha,\beta}$ where:

1. $(1-\kappa)\|\boldsymbol{u}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{u}\|_2^2 \leq (1+\kappa)\|\boldsymbol{u}\|_2^2$,
2. For all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$,
$$\|\boldsymbol{u} - (\boldsymbol{x}^+ - \boldsymbol{x}^-)\|_2 - \alpha \leq \|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^-)\|_2 \leq \|\boldsymbol{u} - (\boldsymbol{x}^+ - \boldsymbol{x}^-)\|_2 + \alpha,$$
3. For all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$,
$$\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 - \beta \leq \|\boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^-)\|_2 \leq \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 + \beta,$$
4. There exists $\boldsymbol{x}^- \in \mathcal{X}^-$ such that $\|\boldsymbol{A}\boldsymbol{x}^-\|_2^2 \leq (1+\kappa)\|\boldsymbol{x}^-\|_2^2$,
5. There exists $\boldsymbol{x}^+ \in \mathcal{X}^+$ such that $\|\boldsymbol{A}\boldsymbol{x}^+\|_2^2 \leq (1+\kappa)\|\boldsymbol{x}^+\|_2^2$.

On the event $\mathsf{B}_{\kappa,\alpha,\beta}$ we clearly have $\boldsymbol{A}\mathcal{X}^-, \boldsymbol{A}\mathcal{X}^+ \subset 5\mathbb{B}_2^k$ and $\|\boldsymbol{A}\boldsymbol{x}^+ - \boldsymbol{A}\boldsymbol{x}^-\|_2 \geq \frac{\gamma}{2}$ for all $\boldsymbol{x}^+ \in \mathcal{X}^+$, $\boldsymbol{x}^- \in \mathcal{X}^-$. Further, by applying Lemma 17 for $\mathcal{X} = \mathcal{X}^+ - \mathcal{X}^- \subset 2\mathbb{B}_2^d$ and $t = 1-\varepsilon$, we obtain that on the event $\mathsf{B}_{\kappa,\alpha,\beta}$,

$$\langle \tfrac{\boldsymbol{A}\boldsymbol{u}}{\|\boldsymbol{A}\boldsymbol{u}\|_2}, \boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^-)\rangle \geq \sqrt{\tfrac{2}{3}}(1-\varepsilon)\|\boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^-)\|_2 - \tfrac{\kappa}{\sqrt{2}} - \sqrt{2}(3+(1-\varepsilon))\beta - \tfrac{3}{\sqrt{2}}(1+2)\alpha$$

for all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$. For $\kappa \lesssim \gamma(1-\varepsilon)$ and $\alpha, \beta \lesssim \gamma(1-\varepsilon)$, we obtain

$$\langle \tfrac{\boldsymbol{A}\boldsymbol{u}}{\|\boldsymbol{A}\boldsymbol{u}\|_2}, \boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^-)\rangle \geq \tfrac{1}{2}(1-\varepsilon)\|\boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^-)\|_2 \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-.$$

Hence, if $\kappa \lesssim \gamma(1-\varepsilon)$ and $\alpha, \beta \lesssim \gamma(1-\varepsilon)$, then $\mathsf{B}_{\kappa,\alpha,\beta} \subset \mathsf{A}_{\varepsilon',\gamma'}$ for $\varepsilon' = \frac{1+\varepsilon}{2}$ and $\gamma' = \frac{\gamma}{2}$. For this choice of ε' and γ', Corollary 15 implies that if $\lambda \geq 5\sqrt{k}$, then

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{v}',\tau}(H[\sqrt{k}\boldsymbol{v}',\tau] \; c\gamma(1-\varepsilon)\sqrt{k}\text{-separates } \boldsymbol{A}\mathcal{X}^- \text{ from } \boldsymbol{A}\mathcal{X}^+ \mid \mathsf{A}_{\varepsilon',\gamma'})\\ &\geq c\tfrac{\gamma(1-\varepsilon)}{\lambda}\sqrt{k}\exp(-Ck\log(4(1-\varepsilon)^{-1})).\end{aligned}$$

Therefore, applying (27) with $\rho = c\gamma(1-\varepsilon)\sqrt{k}$ and $\varepsilon' = \frac{1+\varepsilon}{2}$, $\gamma' = \frac{\gamma}{2}$, we obtain that if $\lambda \geq 5\sqrt{k}$, then

$$\begin{aligned}&\mathbb{P}(H[\boldsymbol{g},\tau] \; c\gamma(1-\varepsilon)\sqrt{k}\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+) \qquad (28)\\ &\geq c\tfrac{\gamma(1-\varepsilon)}{\lambda}\sqrt{k}\exp(-Ck\log(4(1-\varepsilon)^{-1}))\cdot\mathbb{P}(\mathsf{A}_{\varepsilon',\gamma'})\\ &\geq c\tfrac{\gamma(1-\varepsilon)}{\lambda}\sqrt{k}\exp(-Ck\log(4(1-\varepsilon)^{-1}))\cdot\mathbb{P}(\mathsf{B}_{\kappa,\alpha,\beta}),\end{aligned}$$

where the second inequality holds for $\kappa \lesssim \gamma(1-\varepsilon)$ and $\alpha, \beta \lesssim \gamma(1-\varepsilon)$.

Let $\mathcal{T} \subset \mathbb{R}^d$ be a set. By matrix deviation inequality for Gaussian matrices (e.g., see Vershynin, 2018, Sec. 9.1), if

$$k \gtrsim \theta^{-2}(w^2(\mathcal{T}) + \log(2/\eta)\operatorname{rad}^2(\mathcal{T})),$$

then with probability at least $1-\eta$,

$$\sup_{\boldsymbol{x}\in\mathcal{T}} \big|\|\boldsymbol{A}\boldsymbol{x}\|_2 - \|\boldsymbol{x}\|_2\big| \leq \theta.$$

Hence, a union bound implies that if

$$k \gtrsim \kappa^{-2}\log(2/\eta), \quad k \gtrsim (\alpha^{-2}+\beta^{-2})(w^2(\mathcal{X}^+ - \mathcal{X}^-) + \log(2/\eta)),$$

then $\mathsf{B}_{\kappa,\alpha,\beta}$ occurs with probability at least $1-\eta$. In particular, if

$$k \gtrsim \gamma^{-2}(1-\varepsilon)^{-2}(w^2(\mathcal{X}^+ - \mathcal{X}^-) + 1),$$

then $\mathsf{B}_{\kappa,\alpha,\beta}$ with $\kappa \sim \gamma(1-\varepsilon)$ and $\alpha, \beta \sim \gamma(1-\varepsilon)$ occurs with probability at least $\frac{1}{2}$. Combining this result with (28), we obtain that if

$$k \gtrsim \gamma^{-2}(1-\varepsilon)^{-2}(w^2(\mathcal{X}^+ - \mathcal{X}^-) + 1), \quad \lambda \geq 5\sqrt{k},$$

then

$$\mathbb{P}(H[\boldsymbol{g},\tau] \; c\gamma(1-\varepsilon)\sqrt{k}\text{-separates } \mathcal{X}^- \text{ from } \mathcal{X}^+) \geq c\tfrac{\gamma(1-\varepsilon)}{2\lambda}\sqrt{k}\exp(-Ck\log(4(1-\varepsilon)^{-1})).$$

Let $\mu \coloneqq \gamma(1-\varepsilon)$. Setting $k = \mu^{-2}t^2$ completes the proof. ∎

3.2 Separation of Two Points

The following result concerns the separation of two arbitrary points by a random hyperplane, and does not follow directly from Theorem 6. It can be shown in a more elementary way, leading to a stronger statement.

Theorem 18 *There exist absolute constants $c, C > 0$ such that the following holds.*

Let $\boldsymbol{x}^-, \boldsymbol{x}^+ \in R\mathbb{B}_2^d$. Let $\boldsymbol{g} \in \mathbb{R}^d$ denote a standard Gaussian random vector and let $\tau \in [-\lambda, \lambda]$ be uniformly distributed. If $\lambda \geq CR$, then with probability at least $c\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2/\lambda$, the hyperplane $H[\boldsymbol{g}, \tau]$ $\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$-separates $\boldsymbol{x}^-$ from $\boldsymbol{x}^+$.

Proof Since $\frac{\boldsymbol{x}^+ - \boldsymbol{x}^-}{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2} \in \mathbb{S}^{d-1}$, the random variable $\langle \boldsymbol{g}, \frac{\boldsymbol{x}^+ - \boldsymbol{x}^-}{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2} \rangle$ is standard Gaussian. Therefore,

$$\mathbb{P}(\langle \boldsymbol{g}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq 4\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2) \geq c$$

for an absolute constant $c > 0$. Further, we have the inequalities

$$\mathbb{P}(\langle \boldsymbol{g}, \boldsymbol{x}^+ \rangle \leq \lambda) \geq 1 - \exp(-\lambda^2/2\|\boldsymbol{x}^+\|_2^2) \geq 1 - \exp(-\lambda^2/2R^2)$$

and

$$\mathbb{P}(\langle \boldsymbol{g}, \boldsymbol{x}^- \rangle \geq -\lambda) \geq 1 - \exp(-\lambda^2/2R^2).$$

Define the event

$$\mathsf{A} \coloneqq \{\langle \boldsymbol{g}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq 4\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2\} \cap \{\langle \boldsymbol{g}, \boldsymbol{x}^+ \rangle \leq \lambda\} \cap \{\langle \boldsymbol{g}, \boldsymbol{x}^- \rangle \geq -\lambda\}.$$

By the above, $\mathbb{P}(\mathsf{A}) \geq c - 2\exp(-\lambda^2/2R^2)$. Therefore, if $\lambda \geq CR$ for $C > 0$ an absolute constant that is chosen large enough, then $\mathbb{P}(\mathsf{A}) \geq \frac{c}{2}$. Let us show that

$$\mathbb{P}_\tau(H[\boldsymbol{g}, \tau] \ \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2\text{-separates } \boldsymbol{x}^- \text{ from } \boldsymbol{x}^+ \mid \mathsf{A}) \geq \tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2}{\lambda}.$$

Indeed, on the event A it holds $\langle \boldsymbol{g}, \boldsymbol{x}^+ \rangle, \langle \boldsymbol{g}, \boldsymbol{x}^- \rangle \in [-\lambda, \lambda]$ and $\langle \boldsymbol{g}, \boldsymbol{x}^+ - \boldsymbol{x}^- \rangle \geq 4\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$. In particular, the interval $\mathcal{I} \coloneqq [\langle \boldsymbol{g}, \boldsymbol{x}^- \rangle + \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2, \langle \boldsymbol{g}, \boldsymbol{x}^+ \rangle - \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2]$ belongs to $[-\lambda, \lambda]$ with $|\mathcal{I}| \geq 2\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$. If $-\tau \in \mathcal{I}$, then $H[\boldsymbol{g}, \tau]$ $\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$-separates $\boldsymbol{x}^-$ from $\boldsymbol{x}^+$. Therefore,

$$\mathbb{P}_\tau(H[\boldsymbol{g}, \tau] \ \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2\text{-separates } \boldsymbol{x}^- \text{ from } \boldsymbol{x}^+ \mid \mathsf{A}) \geq \mathbb{P}_\tau(-\tau \in \mathcal{I} \mid \mathsf{A}) = \tfrac{|\mathcal{I}|}{2\lambda} \geq \tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2}{\lambda}.$$

The result now follows from

$$\begin{aligned}
&\mathbb{P}(H[\boldsymbol{g}, \tau] \ \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2\text{-separates } \boldsymbol{x}^- \text{ from } \boldsymbol{x}^+) \\
&\geq \mathbb{P}(\{H[\boldsymbol{g}, \tau] \ \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2\text{-separates } \boldsymbol{x}^- \text{ from } \boldsymbol{x}^+\} \cap \mathsf{A}) \\
&= \mathbb{P}(H[\boldsymbol{g}, \tau] \ \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2\text{-separates } \boldsymbol{x}^- \text{ from } \boldsymbol{x}^+ \mid \mathsf{A}) \cdot \mathbb{P}(\mathsf{A}) \\
&\geq \tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2}{\lambda} \cdot \tfrac{c}{2}.
\end{aligned}$$

■

4. Distance Preservation

The following theorem describes how the Euclidean geometry of two sets $\mathcal{X}^-, \mathcal{X}^+$ is transformed by applying a random ReLU-layer. It shows that with high probability, Euclidean distances are approximately preserved provided that both the layer is wide and the bias parameter λ is large enough.

Theorem 19 *There exist absolute constants $C, C', c > 0$ such that the following holds.*

Let $\mathcal{X}^-, \mathcal{X}^+ \subset R\mathbb{B}_2^d$ and let $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ be a random ReLU-layer with maximal bias $\lambda \geq 0$. If $\lambda \geq CR\sqrt{\log(\lambda^2/\varepsilon)}$ for $0 < \varepsilon \leq \lambda^2/e$, and

$$n \geq C'\varepsilon^{-2}\lambda^2\big(w^2(\mathcal{X}^+) + w^2(\mathcal{X}^-) + u^2\lambda^2\big), \tag{29}$$

then with probability at least $1 - 2\exp(-cu^2)$, the following three events occur:

(i) For all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$, we have

$$\Big|\|\Phi(\boldsymbol{x}^+) - \Phi(\boldsymbol{x}^-)\|_2^2 - \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^2\Big(1 - \sqrt{\tfrac{2}{\pi}}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2}{3\lambda}\Big)\Big| \leq \varepsilon. \tag{30}$$

(ii) For all $\boldsymbol{x} \in \mathcal{X}^- \cup \mathcal{X}^+$, we have

$$\Big|\|\Phi(\boldsymbol{x})\|_2^2 - \big(\|\boldsymbol{x}\|_2^2 + \tfrac{\lambda^2}{3}\big)\Big| \leq \varepsilon. \tag{31}$$

(iii) For all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$, we have

$$\Big|\langle\Phi(\boldsymbol{x}^+), \Phi(\boldsymbol{x}^-)\rangle - \Big(\langle\boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^3}{\lambda}\Big)\Big| \leq \varepsilon. \tag{32}$$

The following two results are straightforward corollaries of (31) and (30) in Theorem 19, respectively.

Corollary 20 *There exist absolute constants $C, C' > 0$ such that the following holds.*

Let $\mathcal{X} \subset R\mathbb{B}_2^d$ and let $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ be a random ReLU-layer with maximal bias $\lambda \geq 0$. For any $\eta \in (0,1)$, if $\lambda \geq CR$ and

$$n \geq C'\big(\lambda^{-2}w^2(\mathcal{X}) + \log(e/\eta)\big),$$

then $\Phi(\mathcal{X}) \subset \lambda\mathbb{B}_2^n$ with probability at least $1 - \eta$.

Corollary 21 *There exist absolute constants $C, C' > 0$ such that the following holds.*

Let $\mathcal{X}^-, \mathcal{X}^+ \subset R\mathbb{B}_2^d$ be δ-separated sets and let $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ be a random ReLU-layer with maximal bias $\lambda \geq 0$. For any $\eta \in (0,1)$, if $\lambda \geq CR\sqrt{\log(\lambda/\delta)}$, $\lambda/\delta \geq e$, and

$$n \geq C'\delta^{-4}\lambda^2\big(w^2(\mathcal{X}^+) + w^2(\mathcal{X}^-) + \log(e/\eta)\lambda^2\big),$$

then $\Phi(\mathcal{X}^-)$ and $\Phi(\mathcal{X}^+)$ are $\frac{\delta}{2}$-separated with probability at least $1 - \eta$.

Let us outline the main steps of the proof of Theorem 19. Note that it suffices to show (32). Indeed, (31) trivially follows from (32). Further, (30) follows from (32) and (31) by polarization. To show (32), we proceed in two steps:

1. Compute the expected value of $\langle \Phi(\boldsymbol{x}^+), \Phi(\boldsymbol{x}^-)\rangle$ for two arbitrary points $\boldsymbol{x}^+, \boldsymbol{x}^- \in R\mathbb{B}_2^d$ (see Proposition 22).

2. Show uniform concentration of $\langle \Phi(\boldsymbol{x}^+), \Phi(\boldsymbol{x}^-)\rangle$ around its expected value using a concentration result for empirical product processes due to Mendelson (2016, Thm. 1.13).

Proposition 22 *There exists an absolute constant $C > 0$ such that the following holds.*
Let $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ be a random ReLU-layer with maximal bias $\lambda \geq 0$. If $\lambda \geq CR\sqrt{\log(\lambda^2/\varepsilon)}$ and $\lambda^2/\varepsilon \geq e$, then for any $\boldsymbol{x}^+, \boldsymbol{x}^- \in R\mathbb{B}_2^d$, we have that

$$\left|\mathbb{E}[\langle \Phi(\boldsymbol{x}^+), \Phi(\boldsymbol{x}^-)\rangle] - \left(\langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^3}{\lambda}\right)\right| \leq \varepsilon.$$

To prove Proposition 22, we will make use of the following lemma:

Lemma 23 *Let $\tau \in [-\lambda, \lambda]$ be uniformly distributed. For $a, b \in [-\lambda, \lambda]$,*

$$\mathbb{E}[\mathrm{ReLU}(a+\tau)\,\mathrm{ReLU}(b+\tau)] = \tfrac{ab}{2} + \tfrac{(\min\{a,b\})^2 \max\{a,b\}}{4\lambda} - \tfrac{(\min\{a,b\})^3}{12\lambda} + (a+b)\tfrac{\lambda}{4} + \tfrac{\lambda^2}{6}.$$

Proof We may assume that $a \leq b$. Then

$$\begin{aligned}
\mathbb{E}[\mathrm{ReLU}(a+\tau)\,\mathrm{ReLU}(b+\tau)] &= \mathbb{E}[(a+\tau)\mathbb{1}_{\tau \geq -a}(b+\tau)\mathbb{1}_{\tau \geq -b}] \\
&= \mathbb{E}[(a+\tau)(b+\tau)\mathbb{1}_{\tau \geq -a}] \\
&= \tfrac{1}{2\lambda}\int_{-a}^{\lambda} (ab + (a+b)s + s^2)ds \\
&= \tfrac{1}{2\lambda}\big(ab(\lambda + a) + (a+b)(\tfrac{\lambda^2}{2} - \tfrac{a^2}{2}) + \tfrac{\lambda^3}{3} + \tfrac{a^3}{3}\big) \\
&= \tfrac{ab}{2} + \tfrac{a^2 b}{4\lambda} - \tfrac{a^3}{12\lambda} + (a+b)\tfrac{\lambda}{4} + \tfrac{\lambda^2}{6} \\
&= \tfrac{ab}{2} + \tfrac{(\min\{a,b\})^2 \max\{a,b\}}{4\lambda} - \tfrac{(\min\{a,b\})^3}{12\lambda} + (a+b)\tfrac{\lambda}{4} + \tfrac{\lambda^2}{6}.
\end{aligned}$$

■

Proof [Proposition 22] Clearly, we have that

$$\mathbb{E}[\langle \Phi(\boldsymbol{x}^+), \Phi(\boldsymbol{x}^-)\rangle] = 2 \cdot \mathbb{E}[\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\,\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)],$$

where $\boldsymbol{g}$ denotes a standard Gaussian vector and $\tau \in [-\lambda, \lambda]$ is an independent and uniformly distributed random variable. Let us begin by showing that

$$\begin{aligned}
&\mathbb{E}\Big[\tfrac{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle}{2} + \tfrac{(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^2 \max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\}}{4\lambda} - \tfrac{(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3}{12\lambda} \\
&\quad + \tfrac{\lambda(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle)}{4} + \tfrac{\lambda^2}{6}\Big] \\
&= \tfrac{1}{2}\Big(\langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^3}{\lambda}\Big). \qquad (33)
\end{aligned}$$

Since $\mathbb{E}[\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle] = \langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle$ for any vectors $\boldsymbol{x}^+, \boldsymbol{x}^- \in \mathbb{R}^d$ and we have $\mathbb{E}[\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle] = \mathbb{E}[\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle] = 0$, this amounts to showing that

$$\begin{aligned}&\mathbb{E}\Big[3(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^2 \max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\} - (\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3\Big] \\ &= \sqrt{\tfrac{2}{\pi}}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^3. \end{aligned} \tag{34}$$

Since $\boldsymbol{g}$ is symmetric, it follows that

$$\begin{aligned}&\mathbb{E}\big[(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^2 \max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\}\big] \\ &= -\mathbb{E}\big[(\max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^2 \min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\}\big]\end{aligned}$$

and

$$\mathbb{E}\big[(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3\big] = -\mathbb{E}\big[(\max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3\big].$$

Therefore,

$$\begin{aligned}&\mathbb{E}\Big[3(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^2 \max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\} - (\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3\Big] \\ &= \tfrac{1}{2}\cdot\Big(3\mathbb{E}\big[(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^2 \max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\}\big] \\ &\quad - 3\mathbb{E}\big[(\max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^2 \min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\}\big] \\ &\quad - \mathbb{E}\big[(\min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3\big] + \mathbb{E}\big[(\max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3\big]\Big) \\ &= \tfrac{1}{2}\cdot\mathbb{E}\big[(\max\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\} - \min\{\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle, \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\})^3\big] \\ &= \tfrac{1}{2}\cdot\mathbb{E}\big[|\langle \boldsymbol{g}, \boldsymbol{x}^+ - \boldsymbol{x}^-\rangle|^3\big].\end{aligned}$$

Using that $\boldsymbol{g}$ is rotation invariant and $\mathbb{E}[|g|^3] = 2\sqrt{\frac{2}{\pi}}$ for $g \sim \mathsf{N}(0,1)$, we arrive at (34). Define the event

$$\mathsf{A} = \Big\{ \max\{|\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle|, |\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle|\} \le \lambda\Big\}.$$

Since $\|\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle\|_{\psi_2}, \|\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\|_{\psi_2} \lesssim R$, we have

$$\mathbb{P}(\mathsf{A}^C) \le \mathbb{P}(|\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle| > \lambda) + \mathbb{P}(|\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle| > \lambda) \le 4\exp(-c\lambda^2/R^2)$$

for some absolute constant $c > 0$. By using the Cauchy-Schwarz inequality twice, we obtain

$$\begin{aligned}&\mathbb{E}[|\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)| \cdot \mathbb{1}_{\mathsf{A}^C}] \\ &\le \big(\mathbb{E}\big[(\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau))^2(\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau))^2\big]\big)^{1/2}\cdot \mathbb{P}(\mathsf{A}^C)^{1/2} \\ &\le \|\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\|_{L^4}\cdot\|\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)\|_{L^4}\cdot 2\exp(-c\lambda^2/2R^2) \\ &\lesssim \lambda^2\exp(-c\lambda^2/2R^2) \le \varepsilon,\end{aligned}$$

where the last inequality follows if $\lambda \ge CR\sqrt{\log(\lambda^2/\varepsilon)}$ for $C > 0$ an absolute constant that is chosen large enough and $\lambda^2/\varepsilon \ge e$. Therefore,

$$\begin{aligned}&\Big|\mathbb{E}[\langle \Phi(\boldsymbol{x}^+), \Phi(\boldsymbol{x}^-)\rangle] - \Big(\langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^3}{\lambda}\Big)\Big| \\ &= \Big|2\mathbb{E}[\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)] - \Big(\langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^3}{\lambda}\Big)\Big| \\ &\le \Big|2\mathbb{E}[\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\operatorname{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)\cdot\mathbb{1}_{\mathsf{A}}] - \Big(\langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^3}{\lambda}\Big)\Big| \\ &\quad + 2\varepsilon.\end{aligned} \tag{35}$$

Using Lemma 23 and the independence of $\boldsymbol{g}$ and τ we obtain

$$\begin{aligned}
&\mathbb{E}[\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\,\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)\cdot \mathbb{1}_{\mathsf{A}}]\\
&= \mathbb{E}_{\boldsymbol{g}}[\mathbb{1}_{\mathsf{A}}\mathbb{E}_{\tau}[\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\,\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)]]\\
&= \mathbb{E}_{\boldsymbol{g}}\big[\mathbb{1}_{\mathsf{A}}\big(\tfrac{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle\langle \boldsymbol{g},\boldsymbol{x}^-\rangle}{2} + \tfrac{(\min\{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle,\langle \boldsymbol{g},\boldsymbol{x}^-\rangle\})^2 \max\{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle,\langle \boldsymbol{g},\boldsymbol{x}^-\rangle\}}{4\lambda} - \tfrac{(\min\{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle,\langle \boldsymbol{g},\boldsymbol{x}^-\rangle\})^3}{12\lambda}\big)\big]\\
&\quad + \mathbb{E}_{\boldsymbol{g}}\big[\mathbb{1}_{\mathsf{A}}\big((\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle)\tfrac{\lambda}{4} + \tfrac{\lambda^2}{6}\big)\big].
\end{aligned}$$

In combination with (33) and the Cauchy-Schwarz inequality this yields

$$\begin{aligned}
&\Big|\mathbb{E}[\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\,\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)\cdot \mathbb{1}_{\mathsf{A}}] - \tfrac{1}{2}\Big(\langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+-\boldsymbol{x}^-\|_2^3}{\lambda}\Big)\Big|\\
&\le \mathbb{E}\big[\big|\tfrac{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle\langle \boldsymbol{g},\boldsymbol{x}^-\rangle}{2} + \tfrac{(\min\{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle,\langle \boldsymbol{g},\boldsymbol{x}^-\rangle\})^2 \max\{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle,\langle \boldsymbol{g},\boldsymbol{x}^-\rangle\}}{4\lambda} - \tfrac{(\min\{\langle \boldsymbol{g},\boldsymbol{x}^+\rangle,\langle \boldsymbol{g},\boldsymbol{x}^-\rangle\})^3}{12\lambda}\\
&\quad + (\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \langle \boldsymbol{g}, \boldsymbol{x}^-\rangle)\tfrac{\lambda}{4} + \tfrac{\lambda^2}{6}\big|\cdot \mathbb{1}_{\mathsf{A}^C}\big]\\
&\lesssim \Big(\|\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle\|_{L^4}\cdot\|\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\|_{L^4} + \tfrac{1}{\lambda}\big(\|\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle\|_{L^6}^3 + \|\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\|_{L^6}^3\big)\\
&\quad + \lambda\big(\|\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle\|_{L^2} + \|\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\|_{L^2}\big) + \lambda^2\Big)\cdot \mathbb{P}(\mathsf{A}^C)^{1/2}\\
&\lesssim \lambda^2\cdot\exp(-c\lambda^2/2R^2) \le \varepsilon,
\end{aligned}$$

where the last two inequalities follow by using $\|\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle\|_{\psi_2}, \|\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle\|_{\psi_2} \lesssim R$ and $\lambda \ge CR\sqrt{\log(\lambda^2/\varepsilon)}$ for $C>0$ an absolute constant that is chosen large enough and $\lambda^2/\varepsilon \ge e$. Together with (35) this implies

$$\Big|\mathbb{E}[\langle \Phi(\boldsymbol{x}^+), \Phi(\boldsymbol{x}^-)\rangle] - \Big(\langle \boldsymbol{x}^+, \boldsymbol{x}^-\rangle + \tfrac{\lambda^2}{3} + \sqrt{\tfrac{2}{\pi}}\tfrac{1}{6}\tfrac{\|\boldsymbol{x}^+-\boldsymbol{x}^-\|_2^3}{\lambda}\Big)\Big| \le 4\varepsilon.$$

By rescaling ε, we obtain the result. ∎

Proof [Theorem 19] Let us start by showing (32). Let $\boldsymbol{g} \in \mathbb{R}^d$ be a standard Gaussian random vector and $\tau \in [-\lambda, \lambda]$ be uniformly distributed. Since the ReLU is 1-Lipschitz, it follows that

$$\|\,\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau) - \mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)\|_{\psi_2} \lesssim \|\langle \boldsymbol{g}, \boldsymbol{x}^+ - \boldsymbol{x}^-\rangle\|_{\psi_2} \lesssim \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$$

and

$$\|\,\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}\rangle + \tau)\|_{\psi_2} \lesssim \|\boldsymbol{x}\|_2 + \lambda \le 2\lambda$$

for all $\boldsymbol{x}^+, \boldsymbol{x}^- \in \mathbb{R}^d$ and $\boldsymbol{x} \in R\mathbb{B}_2^d$. Hence, the stochastic processes

$$\Big\{\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^+\rangle + \tau)\Big\}_{\boldsymbol{x}^+\in\mathcal{X}^+}, \quad \Big\{\mathrm{ReLU}(\langle \boldsymbol{g}, \boldsymbol{x}^-\rangle + \tau)\Big\}_{\boldsymbol{x}^-\in\mathcal{X}^-}$$

are sub-Gaussian with respect to the Euclidean metric and their radii in sub-Gaussian norm are bounded by λ. By a concentration result for empirical product processes where each process is sub-Gaussian (Mendelson, 2016, Thm. 1.13), we have

$$\begin{aligned}
&\tfrac{1}{n}\Big|\sum_{i=1}^n \mathrm{ReLU}(\langle \boldsymbol{w}_i, \boldsymbol{x}^+\rangle + b_i)\,\mathrm{ReLU}(\langle \boldsymbol{w}_i, \boldsymbol{x}^-\rangle + b_i)\\
&\qquad\qquad - \mathbb{E}[\mathrm{ReLU}(\langle \boldsymbol{w}_i, \boldsymbol{x}^+\rangle + b_i)\,\mathrm{ReLU}(\langle \boldsymbol{w}_i, \boldsymbol{x}^-\rangle + b_i)]\Big|\\
&\le C\cdot\left(\frac{(w(\mathcal{X}^+) + u\lambda)(w(\mathcal{X}^-) + u\lambda)}{n} + \frac{\lambda(w(\mathcal{X}^+) + w(\mathcal{X}^-)) + u\lambda^2}{\sqrt{n}}\right) \qquad (36)
\end{aligned}$$

uniformly for all $\boldsymbol{x}^+ \in \mathcal{X}^+, \boldsymbol{x}^- \in \mathcal{X}^-$ with probability at least $1 - 2\exp(-cu^2)$. The condition on n given by (29) now implies that the right hand side of (36) is bounded by ε. In combination with Proposition 22 this shows (32) by using the triangle inequality. Analogously, we can show that (31) holds. Finally, by polarization, (32) and (31) imply (30). ■

5. Proof of the Main Result (Theorem 10)

The following lemma and especially its Corollary 25 are crucial ingredients for the proof of Theorem 10. In short, they make the following geometric statement precise: Let $\mathcal{X}^+, \mathcal{X}^- \in \mathbb{R}^d$ be two sets and $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ a deterministic ReLU-layer. If for every $\boldsymbol{x}^+ \in \mathcal{X}^+$ there exists at least one "neuron" that separates $\mathcal{X}^-$ from $\boldsymbol{x}^+$, then the transformed sets $\Phi(\mathcal{X}^-)$ and $\Phi(\mathcal{X}^+)$ are linearly separable.

Lemma 24 *Let $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$. For $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_n]^\mathsf{T} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{b} = (b_1, \dots, b_n) \in \mathbb{R}^n$, define the associated (deterministic) ReLU-layer $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ by*

$$\Phi(\boldsymbol{x}) \coloneqq \sqrt{\tfrac{2}{n}} \cdot \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

Set

$$I \coloneqq \big\{ i \in [n] \mid \langle \boldsymbol{w}_i, \boldsymbol{x}^- \rangle + b_i \le 0 \text{ for all } \boldsymbol{x}^- \in \mathcal{X}^- \big\}$$

and for $\boldsymbol{x}^+ \in \mathcal{X}^+$, define $I_{\boldsymbol{x}^+}(0) \subset [n]$ to be the set of all indices $i \in [n]$ such that $H[\boldsymbol{w}_i, b_i]$ separates $\mathcal{X}^-$ from $\boldsymbol{x}^+$. Assume that $\min_{\boldsymbol{x}^+\in\mathcal{X}^+} |I_{\boldsymbol{x}^+}(0)| \ge 1$. Then $|I| \ge 1$ and the hyperplane $H[\boldsymbol{u}, 0]$ given by the vector $\boldsymbol{u} \in \mathbb{S}_+^{n-1}$ with

$$u_i = \begin{cases} \frac{1}{\sqrt{|I|}}, & i \in I, \\ 0, & \text{otherwise,} \end{cases}$$

separates $\Phi(\mathcal{X}^-)$ from $\Phi(\mathcal{X}^+)$. More precisely,

$$\begin{aligned} &\langle \boldsymbol{u}, \Phi(\boldsymbol{x}^-) \rangle \le 0 \qquad \text{for all } \boldsymbol{x}^- \in \mathcal{X}^-, \\ &\langle \boldsymbol{u}, \Phi(\boldsymbol{x}^+) \rangle \ge \tfrac{1}{n} \sum_{i \in I_{\boldsymbol{x}^+}(0)} |\langle \boldsymbol{w}_i, \boldsymbol{x}^+ \rangle + b_i| \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+. \end{aligned}$$

Proof Since $I_{\boldsymbol{x}^+}(0) \subset I$ for every $\boldsymbol{x}^+ \in \mathcal{X}^+$ and $\min_{\boldsymbol{x}^+\in\mathcal{X}^+} |I_{\boldsymbol{x}^+}(0)| \ge 1$, it follows $|I| \ge 1$. Further, for any $\boldsymbol{x}^- \in \mathcal{X}^-$, we have that

$$\langle \boldsymbol{u}, \Phi(\boldsymbol{x}^-) \rangle = \sum_{i\in I} \tfrac{1}{\sqrt{|I|}} \cdot \sqrt{\tfrac{2}{n}}\, \mathrm{ReLU}(\underbrace{\langle \boldsymbol{w}_i, \boldsymbol{x}^- \rangle + b_i}_{\le 0}) = 0.$$

On the other hand, for any $\boldsymbol{x}^+ \in \mathcal{X}^+$, we have that

$$\langle \boldsymbol{u}, \Phi(\boldsymbol{x}^+) \rangle \ge \sum_{i\in I_{\boldsymbol{x}^+}(0)} \tfrac{1}{\sqrt{|I|}} \cdot \sqrt{\tfrac{2}{n}}\, \mathrm{ReLU}(\langle \boldsymbol{w}_i, \boldsymbol{x}^+ \rangle + b_i) \ge \tfrac{1}{n} \sum_{i\in I_{\boldsymbol{x}^+}(0)} |\langle \boldsymbol{w}_i, \boldsymbol{x}^+ \rangle + b_i|.$$

■

Corollary 25 *Let $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^d$ and $t \geq 0$. For $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_n]^\mathsf{T} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{b} = (b_1, \dots, b_n) \in \mathbb{R}^n$, define the associated (deterministic) ReLU-layer $\Phi \colon \mathbb{R}^d \to \mathbb{R}^n$ by*

$$\Phi(\boldsymbol{x}) \coloneqq \sqrt{\tfrac{2}{n}} \cdot \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

For $\boldsymbol{x}^+ \in \mathcal{X}^+$, define $I_{\boldsymbol{x}^+}(t) \subset [n]$ to be the set of all indices $i \in [n]$ such that $H[\boldsymbol{w}_i, b_i]$ t-separates $\mathcal{X}^-$ from $\boldsymbol{x}^+$. Assume that $\min_{\boldsymbol{x}^+ \in \mathcal{X}^+} |I_{\boldsymbol{x}^+}(t)| \geq n'$ for some $n' \geq 1$. Then $\Phi(\mathcal{X}^-)$ and $\Phi(\mathcal{X}^+)$ are linearly separable with margin $\frac{tn'}{2n}$.

Proof Set

$$I \coloneqq \big\{ i \in [n] \mid \langle \boldsymbol{w}_i, \boldsymbol{x}^- \rangle + b_i \leq 0 \text{ for all } \boldsymbol{x}^- \in \mathcal{X}^- \big\}.$$

By Lemma 24, we have $|I| \geq 1$ and the hyperplane $H[\boldsymbol{u}, 0]$ given by the vector $\boldsymbol{u} \in \mathbb{S}_+^{n-1}$ with

$$u_i = \begin{cases} \frac{1}{\sqrt{|I|}}, & i \in I, \\ 0, & \text{otherwise}, \end{cases}$$

satisfies

$$\begin{aligned} \langle \boldsymbol{u}, \Phi(\boldsymbol{x}^-) \rangle &\leq 0 \qquad \text{for all } \boldsymbol{x}^- \in \mathcal{X}^-, \\ \langle \boldsymbol{u}, \Phi(\boldsymbol{x}^+) \rangle &\geq \tfrac{1}{n} \sum_{i \in I_{\boldsymbol{x}^+}(0)} |\langle \boldsymbol{w}_i, \boldsymbol{x}^+ \rangle + b_i| \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+, \end{aligned}$$

where $I_{\boldsymbol{x}^+}(0)$ is the set of all indices $i \in [n]$ such that $H[\boldsymbol{w}_i, b_i]$ separates $\mathcal{X}^-$ from $\boldsymbol{x}^+$. Clearly, $I_{\boldsymbol{x}^+}(t) \subset I_{\boldsymbol{x}^+}(0)$, which implies for any $\boldsymbol{x}^+ \in \mathcal{X}^+$ that

$$\langle \boldsymbol{u}, \Phi(\boldsymbol{x}^+) \rangle \geq \tfrac{1}{n} \sum_{i \in I_{\boldsymbol{x}^+}(t)} |\langle \boldsymbol{w}_i, \boldsymbol{x}^+ \rangle + b_i| > \tfrac{tn'}{n}.$$

It follows that the hyperplane $H[\boldsymbol{u}, -\frac{tn'}{2n}]$ separates $\Phi(\mathcal{X}^-)$ and $\Phi(\mathcal{X}^+)$ with margin $\frac{tn'}{2n}$. ■

In line with our proof sketch in Section 1.4, the following two results describe the effect of the first random ReLU-layer Φ in the setup of Theorem 10.

Theorem 26 *There exists an absolute constant $c > 0$ such that the following holds.*

Let $\mathcal{X}^-, \mathcal{X}^+ \subset R\mathbb{B}_2^d$ be δ-separated sets with $N^- \coloneqq |\mathcal{X}^-|$, $N^+ \coloneqq |\mathcal{X}^+|$. Let $\Phi \colon \mathbb{R}^d \to \mathbb{R}^n$ be a random ReLU-layer with maximal bias $\lambda \geq 0$. Suppose that $\lambda \gtrsim R$ and

$$n \gtrsim \delta^{-1} \lambda \cdot \log(2N^- N^+ / \eta). \tag{37}$$

Then with probability at least $1 - \eta$, the following event occurs: for every $\boldsymbol{x}^- \in \mathcal{X}^-$ there exists a vector $\boldsymbol{u}_{\boldsymbol{x}^-} \in \mathbb{S}_+^{n-1}$ such that the hyperplane $H[\boldsymbol{u}_{\boldsymbol{x}^-}, 0]$ linearly separates $\Phi(\boldsymbol{x}^-)$ from $\Phi(\mathcal{X}^+)$. The vector $\boldsymbol{u}_{\boldsymbol{x}^-}$ is given by $\boldsymbol{u}_{\boldsymbol{x}^-} = \|\boldsymbol{u}'_{\boldsymbol{x}^-}\|_2^{-1} \boldsymbol{u}'_{\boldsymbol{x}^-}$ for

$$(\boldsymbol{u}'_{\boldsymbol{x}^-})_i = \begin{cases} 1, & (\Phi(\boldsymbol{x}^-))_i = 0, \\ 0, & \textit{otherwise}, \end{cases} \tag{38}$$

and satisfies

$$\langle \boldsymbol{u}_{\boldsymbol{x}^-}, \Phi(\boldsymbol{x}^-)\rangle \le 0,$$
$$\langle \boldsymbol{u}_{\boldsymbol{x}^-}, \Phi(\boldsymbol{x}^+)\rangle \ge c\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2^2 \cdot \lambda^{-1} \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+.$$

Proof Let $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_n]^\mathsf{T} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{b} = (b_1, \dots, b_n) \in \mathbb{R}^n$ be the weight matrix and bias vector of Φ, respectively. For $\boldsymbol{x}^- \in \mathcal{X}^-$, $\boldsymbol{x}^+ \in \mathcal{X}^+$ define $I_{\boldsymbol{x}^-,\boldsymbol{x}^+} \subset [n]$ to be the set of all indices $i \in [n]$ where $H[\boldsymbol{w}_i, b_i]$ separates $\boldsymbol{x}^-$ from $\boldsymbol{x}^+$ and define the events

$$\mathsf{B}^i_{\boldsymbol{x}^-,\boldsymbol{x}^+} \coloneqq \big\{H[\boldsymbol{w}_i, b_i] \ \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2\text{-separates } \boldsymbol{x}^- \text{ from } \boldsymbol{x}^+\big\}.$$

For $n'(\boldsymbol{x}^-, \boldsymbol{x}^+) \in \{1, \dots, n\}$ a number that is specified later, set

$$\mathsf{B}_{\boldsymbol{x}^-,\boldsymbol{x}^+,n'(\boldsymbol{x}^-,\boldsymbol{x}^+)} \coloneqq \Big\{ \sum_{i=1}^n \mathbb{1}_{\mathsf{B}^i_{\boldsymbol{x}^-,\boldsymbol{x}^+}} \ge n'(\boldsymbol{x}^-, \boldsymbol{x}^+)\Big\},$$
$$\mathsf{B}_{\boldsymbol{x}^-} \coloneqq \bigcap_{\boldsymbol{x}^+\in\mathcal{X}^+} \mathsf{B}_{\boldsymbol{x}^-,\boldsymbol{x}^+,n'(\boldsymbol{x}^-,\boldsymbol{x}^+)}, \quad \mathsf{B} \coloneqq \bigcap_{\boldsymbol{x}^-\in\mathcal{X}^-} \mathsf{B}_{\boldsymbol{x}^-}.$$

On the event B, the following holds for every $\boldsymbol{x}^- \in \mathcal{X}^-$: For all $\boldsymbol{x}^+ \in \mathcal{X}^+$ there exist at least $n'(\boldsymbol{x}^-, \boldsymbol{x}^+) \ge 1$ hyperplanes $H[\boldsymbol{w}_i, b_i]$ which $\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$-separate $\boldsymbol{x}^-$ from $\boldsymbol{x}^+$. By Lemma 24, this implies that the following holds on the event B: For every $\boldsymbol{x}^- \in \mathcal{X}^-$ there exists $\boldsymbol{u}_{\boldsymbol{x}^-} \in \mathbb{S}^{n-1}_+$ such that the hyperplane $H[\boldsymbol{u}_{\boldsymbol{x}^-}, 0]$ linearly separates $\Phi(\boldsymbol{x}^-)$ from $\Phi(\mathcal{X}^+)$. More precisely,

$$\langle \boldsymbol{u}_{\boldsymbol{x}^-}, \Phi(\boldsymbol{x}^-)\rangle \le 0,$$
$$\langle \boldsymbol{u}_{\boldsymbol{x}^-}, \Phi(\boldsymbol{x}^+)\rangle \ge \tfrac{1}{n} \sum_{i\in I_{\boldsymbol{x}^-,\boldsymbol{x}^+}} |\langle \boldsymbol{w}_i, \boldsymbol{x}^+\rangle + b_i| \ge \tfrac{n'(\boldsymbol{x}^-,\boldsymbol{x}^+)}{n}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+.$$

Further, Lemma 24 shows that $\boldsymbol{u}_{\boldsymbol{x}^-} = \|\boldsymbol{u}'_{\boldsymbol{x}^-}\|_2^{-1}\boldsymbol{u}'_{\boldsymbol{x}^-}$ for $\boldsymbol{u}'_{\boldsymbol{x}^-}$ with

$$(\boldsymbol{u}'_{\boldsymbol{x}^-})_i = \begin{cases} 1, & \langle \boldsymbol{w}_i, \boldsymbol{x}^-\rangle + b_i \le 0, \\ 0, & \text{otherwise.}\end{cases}$$

Since $\langle \boldsymbol{w}_i, \boldsymbol{x}^-\rangle + b_i \le 0$ is equivalent to $(\Phi(\boldsymbol{x}^-))_i = 0$, this shows that $\boldsymbol{u}_{\boldsymbol{x}^-}$ is given as described in (38). By the union bound, we obtain

$$\mathbb{P}(\mathsf{B}^C) \le \sum_{\boldsymbol{x}^-\in\mathcal{X}^-,\, \boldsymbol{x}^+\in\mathcal{X}^+} \mathbb{P}(\mathsf{B}^C_{\boldsymbol{x}^-,\boldsymbol{x}^+,n'(\boldsymbol{x}^-,\boldsymbol{x}^+)}).$$

Let $i \in [n]$. Theorem 18 implies that if $\lambda \gtrsim R$, then $\mathbb{P}(\mathsf{B}^i_{\boldsymbol{x}^-,\boldsymbol{x}^+}) \ge c_1\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2$ for some absolute constant $c_1 > 0$. Therefore, the Chernoff bound implies that

$$\mathbb{P}\Big(\sum_{i=1}^n \mathbb{1}_{\mathsf{B}^i_{\boldsymbol{x}^-,\boldsymbol{x}^+}} \ge \tfrac{c_1}{2}\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \cdot n\Big) \ge 1 - \exp(-c'\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \cdot n).$$

Setting $n'(\boldsymbol{x}^-, \boldsymbol{x}^+) = \lfloor \tfrac{c_1}{2}\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \cdot n\rfloor$, we obtain

$$\mathbb{P}(\mathsf{B}^C_{\boldsymbol{x}^-,\boldsymbol{x}^+,n'(\boldsymbol{x}^-,\boldsymbol{x}^+)}) \le \exp(-c'\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \cdot n) \le \exp(-c'\lambda^{-1}\delta n).$$

Hence,

$$\mathbb{P}(\mathsf{B}^C) \leq \sum_{\boldsymbol{x}^- \in \mathcal{X}^-,\, \boldsymbol{x}^+ \in \mathcal{X}^+} \exp(-c'\lambda^{-1}\delta n) \leq \eta,$$

where the last inequality follows from

$$n \gtrsim \delta^{-1}\lambda \cdot \big(\log N^- + \log N^+ + \log(\eta^{-1})\big).$$

Finally, observe that (37) implies that $n'(\boldsymbol{x}^-, \boldsymbol{x}^+) = \lfloor \frac{c_1}{2}\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|_2 \cdot n \rfloor \geq 1$ for all $\boldsymbol{x}^- \in \mathcal{X}^-$, $\boldsymbol{x}^+ \in \mathcal{X}^+$. ■

Theorem 27 *There exist absolute constants $c, c' > 0$ such that the following holds.*

Let $\mathcal{X}^-, \mathcal{X}^+ \subset R\mathbb{B}_2^d$ be δ-separated sets. Let $\lambda > 0$ satisfy $\lambda \gtrsim R\sqrt{\log(\lambda/\delta)}$ and $\lambda/\delta \geq e$. Let $\mathcal{C}^+ = \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\} \subset R\mathbb{B}_2^d$ and $\mathcal{C}^- = \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\} \subset R\mathbb{B}_2^d$ be δ-separated and form a λ/c'-mutual covering for $\mathcal{X}^+$ and $\mathcal{X}^-$ with components $\mathcal{X}_1^+, \dots, \mathcal{X}_{N^+}^+ \subset \mathcal{X}^+$ and $\mathcal{X}_1^-, \dots, \mathcal{X}_{N^-}^- \subset \mathcal{X}^-$.

Let $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ be a random ReLU-layer with maximal bias $\lambda \geq 0$, such that

$$\begin{aligned} n &\gtrsim \lambda^{-2} \cdot \big(w^2(\mathcal{X}^-) + w^2(\mathcal{X}^+)\big) + (\tfrac{\lambda}{\delta})^8 \cdot \log(2N^-N^+/\eta), \\ n &\gtrsim \lambda^6 \cdot \Big(\max_{l \in [N^-]} \big\{ \operatorname{dist}^{-8}(\boldsymbol{c}_l^-, \mathcal{C}^+) \cdot w^2(\mathcal{X}_l^-) \big\} + \max_{j \in [N^+]} \big\{ \operatorname{dist}^{-8}(\boldsymbol{c}_j^+, \mathcal{C}^-) \cdot w^2(\mathcal{X}_j^+) \big\} \Big). \end{aligned} \tag{39}$$

Then with probability at least $1 - \eta$, the following two events occur:

1. *$\Phi(\mathcal{X}^-), \Phi(\mathcal{X}^+) \subset \lambda\mathbb{B}_2^n$;*
2. *For every $l \in [N^-]$ there exists a vector $\boldsymbol{u}_{\boldsymbol{c}_l^-} \in \mathbb{S}_+^{n-1}$ such that*

$$\begin{aligned} &\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^-) \rangle - 4c'\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+) \leq -2c'\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+) \qquad \text{for all } \boldsymbol{x}^- \in \mathcal{X}_l^-, \\ &\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^+) \rangle - 4c'\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+) \geq 2c\lambda^{-1} \|\boldsymbol{c}_l^- - \boldsymbol{x}^+\|_2^2 \qquad \text{for all } \boldsymbol{x}^+ \in \mathcal{X}^+. \end{aligned}$$

   Further, $\|\boldsymbol{c}_l^- - \boldsymbol{x}^+\|_2 \geq \frac{1}{2}\operatorname{dist}(\boldsymbol{c}_l^-, \mathcal{C}^+)$ for every $\boldsymbol{x}^+ \in \mathcal{X}^+$, which implies that the hyperplane

$$H[\boldsymbol{u}_{\boldsymbol{c}_l^-}, -4c'\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)]$$

   linearly separates $\Phi(\mathcal{X}_l^-)$ from $\Phi(\mathcal{X}^+)$ with margin $\min\{2c', \frac{c}{2}\}\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$.

Proof Since $\mathcal{C}^+ = \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\}$ and $\mathcal{C}^- = \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\}$ form a λ/c'-mutual covering for $\mathcal{X}^+$ and $\mathcal{X}^-$, there exist $r_1^+, \dots, r_{N^+}^+ \geq 0$ and $r_1^-, \dots, r_{N^-}^- \geq 0$ such that

1. the sets $\mathcal{X}_j^+ \coloneqq \mathcal{X}^+ \cap \mathbb{B}_2^d(\boldsymbol{c}_j^+, r_j^+)$ for $j \in [N^+]$, and $\mathcal{X}_l^- \coloneqq \mathcal{X}^- \cap \mathbb{B}_2^d(\boldsymbol{c}_l^-, r_l^-)$ for $l \in [N^-]$, cover $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively;
2. $r_j^+ \leq c'\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-)$ for $j \in [N^+]$, and $r_l^- \leq c'\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$ for $l \in [N^-]$.

By Corollary 20, if

$$\lambda \gtrsim R, \quad n \gtrsim \lambda^{-2}\big(w^2(\mathcal{X}^-) + w^2(\mathcal{X}^+)\big) + \log(e/\eta),$$

then $\Phi(\mathcal{X}^-), \Phi(\mathcal{X}^+) \subset \lambda\mathbb{B}_2^n$ with probability at least $1-\eta$. Define A to be the event where for every $l \in [N^-]$ there exists a vector $\boldsymbol{u}_{\boldsymbol{c}_l^-} \in \mathbb{S}_+^{n-1}$ such that

$$\begin{aligned}
&\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{c}_l^-)\rangle \le 0,\\
&\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{c}_j^+)\rangle \ge c\lambda^{-1}\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2^2 \qquad \text{for all } j \in [N^+].
\end{aligned}$$

Applying Theorem 26 to $\mathcal{C}^+$ and $\mathcal{C}^-$, the condition (39) implies $\mathbb{P}(\mathsf{A}) \ge 1-\eta$. Define B to be the event where the following holds:

(i) For all $l \in [N^-]$:

$$\sup_{\boldsymbol{x}^- \in \mathcal{X}_l^-} \left| \|\Phi(\boldsymbol{x}^-) - \Phi(\boldsymbol{c}_l^-)\|_2^2 - \|\boldsymbol{x}^- - \boldsymbol{c}_l^-\|_2^2 \Big(1 - \sqrt{\tfrac{2}{\pi}}\tfrac{\|\boldsymbol{x}^- - \boldsymbol{c}_l^-\|_2}{3\lambda}\Big)\right| \le \big(c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)\big)^2,$$

(ii) For all $j \in [N^+]$:

$$\sup_{\boldsymbol{x}^+ \in \mathcal{X}_j^+} \left| \|\Phi(\boldsymbol{x}^+) - \Phi(\boldsymbol{c}_j^+)\|_2^2 - \|\boldsymbol{x}^+ - \boldsymbol{c}_j^+\|_2^2 \Big(1 - \sqrt{\tfrac{2}{\pi}}\tfrac{\|\boldsymbol{x}^+ - \boldsymbol{c}_j^+\|_2}{3\lambda}\Big)\right| \le \big(c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-)\big)^2.$$

By Theorem 19 and a union bound, the condition (39) implies $\mathbb{P}(\mathsf{B}) \ge 1-\eta$.

Let us show that on the event $\mathsf{A} \cap \mathsf{B}$ the second event from Theorem 27 holds. Let $l \in [N^-]$. For any $\boldsymbol{x}^- \in \mathcal{X}_l^-$, we have that

$$\begin{aligned}
\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^-)\rangle &= \langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{c}_l^-)\rangle + \langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^-) - \Phi(\boldsymbol{c}_l^-)\rangle\\
&\le \|\Phi(\boldsymbol{x}^-) - \Phi(\boldsymbol{c}_l^-)\|_2\\
&\le \|\boldsymbol{x}^- - \boldsymbol{c}_l^-\|_2 + c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)\\
&\le r_l^- + c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)\\
&\le 2c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+). \qquad (40)
\end{aligned}$$

Let $\boldsymbol{x}^+ \in \mathcal{X}^+$. Then there exists $j \in [N^+]$ such that $\boldsymbol{x}^+ \in \mathcal{X}_j^+$. It holds

$$\begin{aligned}
\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^+)\rangle &= \langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{c}_j^+)\rangle + \langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^+) - \Phi(\boldsymbol{c}_j^+)\rangle\\
&\ge c\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1} - \|\Phi(\boldsymbol{x}^+) - \Phi(\boldsymbol{c}_j^+)\|_2\\
&\ge c\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1} - \|\boldsymbol{x}^+ - \boldsymbol{c}_j^+\|_2 - c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-)\\
&\ge c\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1} - 2c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-)\\
&\ge c\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1} - 2c'\lambda^{-1}\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2^2\\
&\ge \tfrac{c}{2}\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1},
\end{aligned}$$

where the last inequality follows if $c' \leq \frac{c}{4}$. If $\lambda \gtrsim R$ and $c' \leq 1$, then

$$\begin{aligned}\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2 &\leq \|\boldsymbol{x}^+ - \boldsymbol{c}_j^+\|_2 + \|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2 \\ &\leq c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-) + \|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2 \\ &\leq 2\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2.\end{aligned}$$

Therefore, for any $\boldsymbol{x}^+ \in \mathcal{X}^+$, we obtain

$$\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^+)\rangle \geq \tfrac{c}{8}\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1}. \tag{41}$$

Subtracting $4c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$ in (40) and (41), we obtain that for all $\boldsymbol{x}^- \in \mathcal{X}_l^-$ that

$$\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^-)\rangle - 4c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+) \leq -2c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$$

and for all $\boldsymbol{x}^+ \in \mathcal{X}^+$ that

$$\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^+)\rangle - 4c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+) \geq \tfrac{c}{8}\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1} - 4c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+).$$

If $\lambda \gtrsim R$ and $c' \leq 1$, then for any $\boldsymbol{x}^+ \in \mathcal{X}_j^+$,

$$\begin{aligned}\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2 &\leq \|\boldsymbol{c}_j^+ - \boldsymbol{x}^+\|_2 + \|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2 \\ &\leq c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-) + \|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2 \\ &\leq \tfrac{1}{2}\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2 + \|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2,\end{aligned}$$

which implies $\|\boldsymbol{c}_j^+ - \boldsymbol{c}_l^-\|_2 \leq 2\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2$. In particular, $\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2 \geq \frac{1}{2}\operatorname{dist}(\boldsymbol{c}_l^-, \mathcal{C}^+)$ for all $\boldsymbol{x}^+ \in \mathcal{X}^+$. Furthermore,

$$4c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+) \leq 4c'\lambda^{-1}\|\boldsymbol{c}_l^- - \boldsymbol{c}_j^+\|_2^2 \leq 16c'\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2^2$$

for any $\boldsymbol{x}^+ \in \mathcal{X}_j^+$. Hence, for every $\boldsymbol{x}^+ \in \mathcal{X}^+$, we conclude that

$$\begin{aligned}\langle \boldsymbol{u}_{\boldsymbol{c}_l^-}, \Phi(\boldsymbol{x}^+)\rangle - 4c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+) &\geq \tfrac{c}{8}\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2^2\lambda^{-1} - 16c'\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2^2 \\ &\geq \tfrac{c}{16}\lambda^{-1}\|\boldsymbol{x}^+ - \boldsymbol{c}_l^-\|_2^2,\end{aligned}$$

where the last inequality follows if $c' \leq \frac{c}{256}$. ■

The final ingredient for the proof of Theorem 10 is the following lemma. It provides a sufficient condition under which we have that $w(\Phi(\mathcal{X})) \lesssim w(\mathcal{X})$ with high probability for $\mathcal{X} \subset \mathbb{R}^d$ and $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ a random ReLU-layer.

Lemma 28 *Let $\mathcal{X} \subset \mathbb{R}^d$ and $\Phi\colon \mathbb{R}^d \to \mathbb{R}^n$ be a random ReLU-layer with standard Gaussian weight matrix $\boldsymbol{W} \in \mathbb{R}^{n\times d}$ and maximal bias $\lambda \geq 0$. Then, we have that $w(\Phi(\mathcal{X})) \leq w(\sqrt{\frac{2}{n}}\boldsymbol{W}\mathcal{X})$, and furthermore, the following holds:*

(i) If $n \gtrsim \log(2/\eta)$, then $w(\frac{1}{\sqrt{n}}\boldsymbol{W}\mathcal{X}) \lesssim w(\mathcal{X}) + \sqrt{n}\operatorname{diam}(\mathcal{X})$ with probability at least $1-\eta$.

(ii) If $n \gtrsim w^2(\mathrm{cone}(\mathcal{X}-\mathcal{X}) \cap \mathbb{S}^{d-1}) + \log(2/\eta)$*, then with probability at least* $1-\eta$*,*

$$\sup_{\boldsymbol{x} \neq \boldsymbol{x}' \in \mathcal{X}} \| \tfrac{1}{\sqrt{n}} \boldsymbol{W} (\tfrac{\boldsymbol{x}-\boldsymbol{x}'}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2}) \|_2 \leq 2.$$

On this event $w(\frac{1}{\sqrt{n}}\boldsymbol{W}\mathcal{X}') \leq 2w(\mathcal{X}')$ *and therefore* $w(\Phi(\mathcal{X}')) \leq 2^{3/2} w(\mathcal{X}')$ *for every* $\mathcal{X}' \subset \mathcal{X}$*.*

Proof Let us write $\Phi(\boldsymbol{x}) = \mathrm{ReLU}(\boldsymbol{T}(\boldsymbol{x}))$ for

$$\boldsymbol{T}(\boldsymbol{x}) \coloneqq \sqrt{\tfrac{2}{n}} \boldsymbol{W}\boldsymbol{x} + \sqrt{\tfrac{2}{n}} \boldsymbol{b},$$

where $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ is a standard Gaussian random matrix and $\boldsymbol{b}$ is uniformly distributed on $[-\lambda, \lambda]^n$. Since the ReLU is 1-Lipschitz, the Gaussian version of Talagrand's contraction principle (see, e.g., Vershynin, 2018, Ex. 7.2.13) implies that $w(\Phi(\mathcal{X})) \leq w(\boldsymbol{T}(\mathcal{X}))$. Let $\boldsymbol{g} \in \mathbb{R}^n$ denote a standard Gaussian vector. Since $\mathbb{E}_{\boldsymbol{g}} \langle \boldsymbol{g}, \boldsymbol{x} \rangle = 0$ for every vector $\boldsymbol{x}$, it follows that

$$w(\boldsymbol{T}(\mathcal{X})) = \mathbb{E}_{\boldsymbol{g}} \Big[\sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{g}, \sqrt{\tfrac{2}{n}} \boldsymbol{W}\boldsymbol{x} + \sqrt{\tfrac{2}{n}} \boldsymbol{b} \rangle \Big] = \mathbb{E}_{\boldsymbol{g}} \Big[\sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{g}, \sqrt{\tfrac{2}{n}} \boldsymbol{W}\boldsymbol{x} \rangle \Big].$$

Therefore, $w(\Phi(\mathcal{X})) \leq w(\sqrt{\frac{2}{n}}\boldsymbol{W}\mathcal{X})$. Since $w(\mathcal{S}) \leq \frac{\sqrt{n}}{2} \operatorname{diam}(\mathcal{S})$ for any $\mathcal{S} \subset \mathbb{R}^n$, it follows $w(\frac{1}{\sqrt{n}}\boldsymbol{W}\mathcal{X}) \leq \frac{1}{2} \operatorname{diam}(\boldsymbol{W}\mathcal{X})$. By Gaussian projection (e.g., see Vershynin, 2018, Sec. 7.7), there exists an absolute constant $C > 0$ such that

$$\operatorname{diam}(\boldsymbol{W}\mathcal{X}) \leq C \cdot (w(\mathcal{X}) + \sqrt{n} \operatorname{diam}(\mathcal{X}))$$

with probability at least $1 - 2\exp(-n)$. Define

$$\| \tfrac{1}{\sqrt{n}} \boldsymbol{W} \|_{\mathcal{X}} \coloneqq \sup_{\boldsymbol{x} \neq \boldsymbol{x}' \in \mathcal{X}} \| \tfrac{1}{\sqrt{n}} \boldsymbol{W} (\tfrac{\boldsymbol{x}-\boldsymbol{x}'}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2}) \|_2.$$

Let $\mathcal{X}' \subset \mathcal{X}$. Then $\| \frac{1}{\sqrt{n}} \boldsymbol{W}\boldsymbol{x} - \frac{1}{\sqrt{n}} \boldsymbol{W}\boldsymbol{x}' \|_2 \leq \| \frac{1}{\sqrt{n}} \boldsymbol{W} \|_{\mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}'$, which implies $w(\frac{1}{\sqrt{n}}\boldsymbol{W}\mathcal{X}') \leq \| \frac{1}{\sqrt{n}} \boldsymbol{W} \|_{\mathcal{X}} w(\mathcal{X}')$ by the Sudakov-Fernique inequality. By a Gaussian deviation inequality (e.g., see Vershynin, 2018, Sec. 9.1), if

$$n \gtrsim w^2(\mathrm{cone}(\mathcal{X}-\mathcal{X}) \cap \mathbb{S}^{d-1}) + \log(2/\eta),$$

then $\| \frac{1}{\sqrt{n}} \boldsymbol{W} \|_{\mathcal{X}} \leq 2$ with probability at least $1 - \eta$. ∎

We are now ready to prove the main result of this work:

Proof [Theorem 10] Let $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_n]^\mathsf{T} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{b} = (b_1, \dots, b_n) \in [-\lambda, \lambda]^n$ be the Gaussian weight matrix and bias vector of the random ReLU-layer Φ, and let $\hat{\boldsymbol{W}} = [\hat{\boldsymbol{w}}_1, \dots, \hat{\boldsymbol{w}}_{\hat{n}}]^\mathsf{T} \in \mathbb{R}^{\hat{n} \times n}$ and $\hat{\boldsymbol{b}} = (\hat{b}_1, \dots, \hat{b}_{\hat{n}}) \in [-\hat{\lambda}, \hat{\lambda}]^{\hat{n}}$ be the Gaussian weight matrix and bias vector of the random ReLU-layer $\hat{\Phi}$. Since $\mathcal{X}^+$ and $\mathcal{X}^-$ have $(R, \delta, C'\lambda)$-mutual complexity (N^+, N^-, w^+, w^-), there exists a $C'\lambda$-mutual covering $\mathcal{C}^+ = \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\} \subset \mathbb{R}^d$ and $\mathcal{C}^- = \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\} \subset \mathbb{R}^d$ for $\mathcal{X}^+$ and $\mathcal{X}^-$ such that

(i) $\max_{j\in[N^+]} w(\mathcal{X}_j^+) \le w^+$ and $\max_{l\in[N^-]} w(\mathcal{X}_l^-) \le w^-$;

(ii) $\mathcal{C}^+, \mathcal{C}^- \subset R\mathbb{B}_2^d$ are δ-separated.

Here, $\mathcal{X}_1^+, \dots, \mathcal{X}_{N^+}^+ \subset \mathcal{X}^+$ and $\mathcal{X}_1^-, \dots, \mathcal{X}_{N^-}^- \subset \mathcal{X}^-$ are the components of the covering. Let $C' \coloneqq \frac{1}{c'}$, where $c' > 0$ is the absolute constant from Theorem 27. According to Theorem 27, the condition (15) implies that with probability at least $1-\eta$, the following event A occurs:

1. $\Phi(\mathcal{X}^-), \Phi(\mathcal{X}^+) \subset \lambda\mathbb{B}_2^n$;
2. For every $l \in [N^-]$, there exists a vector $\boldsymbol{u}_{\boldsymbol{c}_l^-} \in \mathbb{S}_+^{n-1}$ such that the hyperplane

$$H[\boldsymbol{u}_{\boldsymbol{c}_l^-}, -4c'\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)]$$

linearly separates $\Phi(\mathcal{X}_l^-)$ from $\Phi(\mathcal{X}^+)$ with margin $\min\{2c', \frac{c}{2}\}\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$.

Here, $c > 0$ is the absolute constant from Theorem 27. By Proposition 13 (iv), and on the event A, the sets $\Phi(\mathcal{X}_l^-)$ and $\Phi(\mathcal{X}^+)$ are contained in $\lambda\mathbb{B}_2^n$ for every $l \in [N^-]$ and they are $(\varepsilon_l, \gamma_l)$-linearly separable with

$$\varepsilon_l = 1 - \min\{2c', \tfrac{c}{2}\}\lambda^{-2}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+), \quad \gamma_l = \min\{4c', c\}\lambda^{-1}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+).$$

For $l \in [N^-]$, $t_l \gtrsim w(\Phi(\mathcal{X}_l^-) - \Phi(\mathcal{X}^+)) + \lambda$, and $i \in [\hat{n}]$, we define the event

$$\mathsf{B}_l^i(t_l) \coloneqq \{H[\hat{\boldsymbol{w}}_i, \hat{b}_i] \text{ } t_l\text{-separates } \Phi(\mathcal{X}^+) \text{ from } \Phi(\mathcal{X}_l^-)\}.$$

Set $\mu_l = \gamma_l(1-\varepsilon_l)$. By Theorem 6, if $\hat{\lambda} \gtrsim \lambda t_l \mu_l^{-1}$, then $\mathbb{P}(\mathsf{B}_l^i(t_l) \mid \mathsf{A}) \ge p_l$ for

$$p_l = \tfrac{t_l}{\hat{\lambda}} \exp(-Ct_l^2\mu_l^{-2}\log(4(1-\varepsilon_l)^{-1})).$$

Define A' to be the event where

$$\sup_{\boldsymbol{x}\neq\boldsymbol{x}'\in\mathcal{X}^-} \|\tfrac{1}{\sqrt{n}}\boldsymbol{W}(\tfrac{\boldsymbol{x}-\boldsymbol{x}'}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2})\|_2 \le 2, \qquad \sup_{\boldsymbol{x}\neq\boldsymbol{x}'\in\mathcal{X}^+} \|\tfrac{1}{\sqrt{n}}\boldsymbol{W}(\tfrac{\boldsymbol{x}-\boldsymbol{x}'}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2})\|_2 \le 2.$$

By Lemma 28 and the union bound, condition (15) implies $\mathbb{P}(\mathsf{A}') \ge 1-\eta$. Further, Lemma 28 shows that on the event A', we have that

$$w(\Phi(\mathcal{X}_l^-) - \Phi(\mathcal{X}^+)) = w(\Phi(\mathcal{X}_l^-)) + w(\Phi(\mathcal{X}^+)) \le 2^{3/2}(w(\mathcal{X}_l^-) + w(\mathcal{X}^+))$$

for every $l \in [N^-]$. Using that $\mu_l \gtrsim \lambda^{-3}\delta^4$ and $1-\varepsilon_l \gtrsim \lambda^{-2}\delta^2$ for every $l \in [N^-]$, we obtain that for every $i \in [\hat{n}]$, $l \in [N^-]$, and $t \asymp w^- + w(\mathcal{X}^+) + \lambda$, if $\hat{\lambda} \gtrsim \lambda^4\delta^{-4}t$, then $\mathbb{P}(\mathsf{B}_l^i(t) \mid \mathsf{A}\cap\mathsf{A}') \ge p(t)$ for

$$p(t) = \tfrac{t}{\hat{\lambda}} \exp(-Ct^2\lambda^6\delta^{-8}\log(\lambda/\delta)).$$

Define the events

$$\mathsf{B}_l(t) \coloneqq \Big\{\sum_{i=1}^{\hat{n}} \mathbb{1}_{\mathsf{B}_l^i(t)} \ge \tfrac{p(t)}{2}\hat{n}\Big\}, \quad \mathsf{B}_t \coloneqq \bigcap_{l\in[N^-]} \mathsf{B}_l(t).$$

By Chernoff's inequality, there exists an absolute constant $c > 0$ such that for all $l \in [N^-]$, it holds that

$$\mathbb{P}(\mathsf{B}_l(t) \mid \mathsf{A} \cap \mathsf{A}') \geq 1 - \exp(-c \cdot p(t)\hat{n}).$$

On the event B_t, for every $l \in [N^-]$ at least $\frac{p(t)}{2}\hat{n}$ out of the $\hat{n}$ hyperplanes $H[\hat{\boldsymbol{w}}_i, \hat{b}_i]$ t-separate $\Phi(\mathcal{X}^+)$ from $\Phi(\mathcal{X}_l^-)$. Using that $\Phi(\mathcal{X}^-) = \bigcup_{l\in[N^-]} \Phi(\mathcal{X}_l^-)$, Corollary 25 implies that $F(\mathcal{X}^+)$ and $F(\mathcal{X}^-)$ are linearly separable with margin

$$\tfrac{tp(t)}{4} \asymp \tfrac{(w^- + w(\mathcal{X}^+) + \lambda)^2}{\hat{\lambda}} \cdot \exp(-C(w^- + w(\mathcal{X}^+) + \lambda)^2 \lambda^6 \delta^{-8} \log(\lambda/\delta)).$$

Define B' to be the event where $F(\mathcal{X}^-), F(\mathcal{X}^+) \subset \hat{\lambda}\mathbb{B}_2^{\hat{n}}$. On the event $\mathsf{B}_t \cap \mathsf{B}'$, the conclusion of Theorem 10 holds. Now, we observe that

$$\begin{aligned}
\mathbb{P}(\mathsf{B}_t \cap \mathsf{B}') &\geq \mathbb{P}(\mathsf{B}_t \cap \mathsf{B}' \cap \mathsf{A} \cap \mathsf{A}') \\
&= \mathbb{P}(\mathsf{B}_t \cap \mathsf{B}' \mid \mathsf{A} \cap \mathsf{A}') \cdot \mathbb{P}(\mathsf{A} \cap \mathsf{A}') \\
&\geq (1 - \mathbb{P}(\mathsf{B}_t^C \mid \mathsf{A} \cap \mathsf{A}') - \mathbb{P}((\mathsf{B}')^C \mid \mathsf{A} \cap \mathsf{A}')) \cdot (1 - \mathbb{P}(\mathsf{A}^C) - \mathbb{P}((\mathsf{A}')^C)) \\
&\geq (1 - \mathbb{P}(\mathsf{B}_t^C \mid \mathsf{A} \cap \mathsf{A}') - \mathbb{P}((\mathsf{B}')^C \mid \mathsf{A} \cap \mathsf{A}')) \cdot (1 - 2\eta).
\end{aligned}$$

The union bound implies

$$\mathbb{P}(\mathsf{B}_t^C \mid \mathsf{A} \cap \mathsf{A}') \leq \sum_{l\in[N^-]} \mathbb{P}((\mathsf{B}_l(t))^C \mid \mathsf{A} \cap \mathsf{A}') \leq \exp(\log(N^-) - cp(t)\hat{n}) \leq \eta,$$

where the last inequality follows from

$$\hat{n} \gtrsim (p(t))^{-1} \log(N^-/\eta).$$

On the event $\mathsf{A}\cap\mathsf{A}'$ it holds $\Phi(\mathcal{X}^-), \Phi(\mathcal{X}^+) \subset \lambda\mathbb{B}_2^n$ and $w(\Phi(\mathcal{X}^-)) \leq 2^{3/2} w(\mathcal{X}^-), w(\Phi(\mathcal{X}^+)) \leq 2^{3/2} w(\mathcal{X}^+)$. Consequently, by Corollary 20, if $\hat{\lambda} \gtrsim \lambda$ and

$$\hat{n} \gtrsim (\hat{\lambda})^{-2}\big(w^2(\mathcal{X}^+) + w^2(\mathcal{X}^-)\big) + \log(e/\eta), \tag{42}$$

then $\mathbb{P}((\mathsf{B}')^C \mid \mathsf{A} \cap \mathsf{A}') \leq \eta$. By Lemma 29,

$$w(\mathcal{X}^-) = w\Big(\bigcup_{l\in[N^-]} \mathcal{X}_l^-\Big) \leq w^- + CR\sqrt{\log N^-}$$

for $C > 0$ an absolute constant. Therefore, condition (16) implies $\hat{\lambda} \gtrsim \lambda$ and (42). In total, we have $\mathbb{P}(\mathsf{B}_t \cap \mathsf{B}') \geq (1 - 2\eta)^2 \geq 1 - 4\eta$. This completes the proof. ∎

6. Proofs of Special-Case Results

To apply our main result, Theorem 10, to various special cases, the following lemma will prove very useful. Although the inequalities stated therein are well-known (e.g., see Jacques and Cambareri, 2017, Lem. 10 for the first inequality), we give a proof for the sake of completeness.

Lemma 29 *There exists an absolute constant $C > 0$ such that the following holds. Let $\mathcal{X}_j \subset R\mathbb{B}_2^d$ for $j \in [N]$. Then*

$$w(\bigcup_{j\in[N]} \mathcal{X}_j) \leq \max_{j\in[N]} w(\mathcal{X}_j) + C \cdot R\sqrt{\log N}. \tag{43}$$

If all sets $\mathcal{X}_j$ additionally satisfy $\operatorname{diam}(\mathcal{X}_j) \leq r$ for some $r > 0$, then

$$w(\bigcup_{j\in[N]} \mathcal{X}_j) \lesssim r\sqrt{d} + R\sqrt{\log N}. \tag{44}$$

Furthermore, if all sets $\mathcal{X}_j$ are finite with $\operatorname{diam}(\mathcal{X}_j) \leq r_j$, then

$$w(\bigcup_{j\in[N]} \mathcal{X}_j) \lesssim \max_{j\in[N]}(r_j\sqrt{\log|\mathcal{X}_j|}) + R\sqrt{\log N}. \tag{45}$$

Proof Let us start by showing (43). Let $\boldsymbol{g} \in \mathbb{R}^d$ denote a standard Gaussian random vector and for $j \in [N]$ pick any $\boldsymbol{c}_j \in \mathcal{X}_j$. Then $\operatorname{rad}(\mathcal{X}_j - \boldsymbol{c}_j) \leq 2R$. Set $X_j \coloneqq \sup_{\boldsymbol{x}\in\mathcal{X}_j}\langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{c}_j\rangle$. Then

$$w(\bigcup_{j\in[N]} \mathcal{X}_j) = \mathbb{E}\max_{j\in[N]}(X_j + \langle \boldsymbol{g}, \boldsymbol{c}_j\rangle) \leq \max_{j\in[N]} \mathbb{E}X_j + \mathbb{E}\max_{j\in[N]}(X_j - \mathbb{E}X_j) + \mathbb{E}\max_{j\in[N]}\langle \boldsymbol{g}, \boldsymbol{c}_j\rangle.$$

Clearly, $\mathbb{E}X_j = w(\mathcal{X}_j)$. By Gaussian Lipschitz concentration (e.g., see Foucart and Rauhut, 2013, Thm. 8.34), we conclude that $X_j - \mathbb{E}X_j$ is a sub-Gaussian random variable with $\|X_j - \mathbb{E}X_j\|_{\psi_2} \lesssim \operatorname{rad}(\mathcal{X}_j - \boldsymbol{c}_j) \leq 2R$. Further, the random variables $\langle \boldsymbol{g}, \boldsymbol{c}_j\rangle$ are sub-Gaussian with $\|\langle \boldsymbol{g}, \boldsymbol{c}_j\rangle\|_{\psi_2} \lesssim \|\boldsymbol{c}_j\|_2 \leq R$. Inequality (43) now follows by applying the maximal inequality for sub-Gaussian random variables (e.g., see Boucheron et al., 2013, Thm. 2.5). Inequalities (44) and (45) immediately follow from (43) by using the standard estimates $w(\mathbb{B}_2^d) \lesssim \sqrt{d}$ and $w(\mathcal{X}) \lesssim \sqrt{\log(|\mathcal{X}|)}$ for any finite $\mathcal{X} \subset \mathbb{B}_2^d$. ■

6.1 Proof of Theorem 2

Define $\mathcal{C}^+ \coloneqq \mathcal{X}^+$ and $\mathcal{C}^- \coloneqq \mathcal{X}^-$. Then $\mathcal{C}^+, \mathcal{C}^- \subset \mathbb{B}_2^d$ are δ-separated. We may write $\mathcal{C}^+ = \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\}$ and $\mathcal{C}^- = \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\}$. Clearly, the sets $\mathcal{X}_j^+ \coloneqq \mathcal{X}^+ \cap \mathbb{B}_2^d(\boldsymbol{c}_j^+, 0) = \{\boldsymbol{c}_j^+\}$ for $j \in [N^+]$ and $\mathcal{X}_l^- \coloneqq \mathcal{X}^- \cap \mathbb{B}_2^d(\boldsymbol{c}_l^-, 0) = \{\boldsymbol{c}_l^-\}$ for $l \in [N^-]$ cover $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively. Let $C' > 0$ denote the absolute constant from Theorem 10. Then

$$0 \leq \tfrac{1}{C'\lambda}\operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-), \quad 0 \leq \tfrac{1}{C'\lambda}\operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$$

for all $j \in [N^+]$ and $l \in [N^-]$. Therefore, $\mathcal{C}^+$ and $\mathcal{C}^-$ form a $C'\lambda$-mutual covering for $\mathcal{X}^+$ and $\mathcal{X}^-$. Moreover, $\mathcal{X}^+$ and $\mathcal{X}^-$ have $(1, \delta, C'\lambda)$-mutual complexity (N^+, N^-, w^+, w^-) with $w^+ = w^- = 0$. Since $w(\mathcal{X}^+) \lesssim \sqrt{\log N^+}$ and

$$w^2(\operatorname{cone}(\mathcal{X}^- - \mathcal{X}^-) \cap \mathbb{S}^{d-1}) + w^2(\operatorname{cone}(\mathcal{X}^+ - \mathcal{X}^+) \cap \mathbb{S}^{d-1}) \lesssim \log N^- + \log N^+,$$

the result follows from Theorem 10. ■

6.2 Proof of Theorem 3

Let $r \leq \frac{1}{C'\lambda}\delta^2$, where $C' > 0$ denotes the absolute constant from Theorem 10. Set $\mathcal{C}^+ := \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\}$ and $\mathcal{C}^- := \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\}$. Then, $\mathcal{C}^+, \mathcal{C}^- \subset \mathbb{B}_2^d$ are δ-separated and the sets $\mathcal{X}_j^+ := \mathcal{X}^+ \cap \mathbb{B}_2^d(\boldsymbol{c}_j^+, r)$ for $j \in [N^+]$, and $\mathcal{X}_l^- := \mathcal{X}^- \cap \mathbb{B}_2^d(\boldsymbol{c}_l^-, r)$ for $l \in [N^-]$, cover $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively. Furthermore, the δ-separability and the assumption $r \lesssim \delta^2/\lambda$ imply that

$$r \leq \tfrac{1}{C'\lambda} \operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-), \quad r \leq \tfrac{1}{C'\lambda} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$$

for all $j \in [N^+], l \in [N^-]$. This shows that $\mathcal{C}^+$ and $\mathcal{C}^-$ form a $C'\lambda$-mutual covering for $\mathcal{X}^+$ and $\mathcal{X}^-$. Therefore, $\mathcal{X}^+$ and $\mathcal{X}^-$ have $(1, \delta, C'\lambda)$-mutual complexity (N^+, N^-, w^+, w^-) with $w^+ = \max_{j\in[N^+]} w(\mathcal{X}_j^+)$ and $w^- = \max_{l\in[N^-]} w(\mathcal{X}_l^-)$. By Lemma 29,

$$w(\mathcal{X}^+) = w\Big(\bigcup_{j\in[N^+]} \mathbb{B}_2^d(\boldsymbol{c}_j^+, r)\Big) \lesssim r\sqrt{d} + \sqrt{\log N^+},$$

and for any $l \in [N^-]$,

$$w(\mathcal{X}_l^-) = w(\mathcal{X}^- \cap \mathbb{B}_2^d(\boldsymbol{c}_l^-, r)) \lesssim r\sqrt{d},$$

which yields $w^- \lesssim r\sqrt{d}$. Analogously, it follows that $w^+ \lesssim r\sqrt{d}$. The result now follows from Theorem 10 by observing that

$$w^2(\operatorname{cone}(\mathcal{X}^- - \mathcal{X}^-) \cap \mathbb{S}^{d-1}) + w^2(\operatorname{cone}(\mathcal{X}^+ - \mathcal{X}^+) \cap \mathbb{S}^{d-1}) \lesssim d.$$

■

6.3 Proof of Theorem 4

For an absolute constant $c > 0$ that is specified later, let $\mathcal{C}^+ = \{\boldsymbol{c}_1^+, \dots, \boldsymbol{c}_{N^+}^+\} \subset \mathcal{X}^+$ and $\mathcal{C}^- = \{\boldsymbol{c}_1^-, \dots, \boldsymbol{c}_{N^-}^-\} \subset \mathcal{X}^-$ be minimal $c\delta^2/\lambda$-coverings of $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively. Then $N^+ = \mathcal{N}(\mathcal{X}^+, c\delta^2/\lambda)$ and $N^- = \mathcal{N}(\mathcal{X}^-, c\delta^2/\lambda)$. Since $\mathcal{X}^+$ and $\mathcal{X}^-$ are δ-separated, it follows that $\mathcal{C}^+$ and $\mathcal{C}^-$ are δ-separated as well. By definition of covering, the sets $\mathcal{X}_j^+ := \mathcal{X}^+ \cap \mathbb{B}_2^d(\boldsymbol{c}_j^+, c\delta^2/\lambda)$ for $j \in [N^+]$, and $\mathcal{X}_l^- := \mathcal{X}^- \cap \mathbb{B}_2^d(\boldsymbol{c}_l^-, c\delta^2/\lambda)$ for $l \in [N^-]$, cover $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively. Further, since $\mathcal{C}^+$ and $\mathcal{C}^-$ are δ-separated, we have that

$$c\delta^2/\lambda \leq c\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_j^+, \mathcal{C}^-), \quad c\delta^2/\lambda \leq c\lambda^{-1} \operatorname{dist}^2(\boldsymbol{c}_l^-, \mathcal{C}^+)$$

for all $j \in [N^+], l \in [N^-]$. This shows that $\mathcal{C}^+$ and $\mathcal{C}^-$ form a $\frac{\lambda}{c}$-mutual covering for $\mathcal{X}^+$ and $\mathcal{X}^-$. Therefore, $\mathcal{X}^+$ and $\mathcal{X}^-$ have $(1, \delta, \frac{\lambda}{c})$-mutual complexity (N^+, N^-, w^+, w^-) with $N^+ = \mathcal{N}(\mathcal{X}^+, c\delta^2/\lambda)$, $N^- = \mathcal{N}(\mathcal{X}^-, c\delta^2/\lambda)$, $w^+ = w(\mathcal{X}^+)$ and $w^- = w(\mathcal{X}^-)$. Choosing $c = \frac{1}{C'}$, where C' is the absolute constant from Theorem 10, the result follows from Theorem 10. ■

Acknowledgments

S.D. and M.G. acknowledge support by the DFG Priority Programme DFG-SPP 1798 Grant DI 2120/1-1. A.S. acknowledges support by the Fonds de la Recherche Scientifique – FNRS under Grant n° T.0136.20 (Learn2Sense). L.J. is a FNRS Senior Research Associate.

References

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 242–252, 2019.

S. An, F. Boussaid, and M. Bennamoun. How can deep rectifier networks achieve linear separability and preserve distances? In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 514–523, 2015.

A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang. Learning polynomials with neural networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML)*, pages 1908–1916, 2014.

S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32*, volume 32, 2019.

D. Arpit and Y. Bengio. The benefits of over-parameterization at initialization in deep relu networks. Preprint arXiv:1901.03611, 2019.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.*, 18(1):714–751, 2017.

A. S. Bandeira, D. G. Mixon, and B. Recht. Compressive classification and the rare eclipse problem. In H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, and R. Mathar, editors, *Compressed Sensing and its Applications: Second International MATHEON Conference 2015*, Applied and Numerical Harmonic Analysis, pages 197–220. Springer Cham, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

G. Bresler and D. Nagaraj. A corrective view of neural networks: Representation, memorization and learning. In *Proceedings of Thirty Third Conference on Learning Theory (COLT)*, pages 848–901, 2020.

V. Cambareri, C. Xu, and L. Jacques. The rare eclipse problem on tiles: Quantised embeddings of disjoint convex sets. In *Proceedings of the 2017 International Conference on Sampling Theory and Applications (SampTA)*, 2017.

Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems 32*, 2019a.

Y. Cao and Q. Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. Preprint arXiv:1902.01384, 2019b.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

Z. Chen, Y. Cao, D. Zou, and Q. Gu. How much over-parameterization is sufficient to learn deep relu networks? *Preprint arXiv:1911.12360*, 2019.

S. Dirksen. Quantized compressed sensing: A survey. In H. Boche, G. Caire, R. Calderbank, G. Kutyniok, R. Mathar, and P. Petersen, editors, *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*, Applied and Numerical Harmonic Analysis, pages 67–95. Birkhäuser Cham, 2019.

S. Dirksen and S. Mendelson. Robust one-bit compressed sensing with partial circulant matrices. *Ann. Appl. Probab., to appear. Preprint arXiv:1812.06719*, 2018.

S. Dirksen and S. Mendelson. Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing. *J. Eur. Math. Soc.*, 23(9):2913–2947, 2021.

S. Dirksen, S. Mendelson, and A. Stollenwerk. Sharp estimates on random hyperplane tessellations. *SIAM J. Math. Data Sci., to appear. Preprint arXiv:2201.05204*, 2022a.

S. Dirksen, S. Mendelson, and A. Stollenwerk. Fast metric embedding into the Hamming cube. Preprint arXiv:2204.04109, 2022b.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1675–1685, 2019.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser Basel, 2013.

A. A. Giannopoulos and V. D. Milman. Asymptotic convex geometry short overview. In S. Donaldson, Y. Eliashberg, and M. Gromov, editors, *Different Faces of Geometry*, pages 87–162. Springer Boston, 2004.

R. Giryes, G. Sapiro, and A. M. Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Trans. Signal Process.*, 64(13):3444–3457, 2016.

R. Giryes, G. Sapiro, and A. M. Bronstein. Corrections to: "deep neural networks with random gaussian weights: A universal classification strategy?". *IEEE Trans. Signal Process.*, 68:529–531, 2020.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In J. Lindenstrauss and V. D. Milman, editors, *Geometric Aspects of Functional Analysis*, volume 1317 of *Lecture Notes in Mathematics*, pages 84–106. Springer Berlin Heidelberg, 1988.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

D. Hsu, C. Sanford, R. A. Servedio, and E.-V. Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. Preprint arXiv:2102.02336, 2021.

G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. Preprint arXiv:1806.07572, 2018.

L. Jacques and V. Cambareri. Time for dithering: fast and quantized random embeddings via the restricted isometry property. *Inf. Inference*, 6(4):441–476, 2017.

L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Trans. Inf. Theory*, 59(4): 2082–2102, 2013.

Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations (ICLR)*, 2020.

H. C. Jung, J. Maly, L. Palzer, and A. Stollenwerk. Quantized compressed sensing by rectified linear units. *IEEE Trans. Inf. Theory*, 67(6):4125–4149, 2021.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems 31*, 2018.

Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic. Towards a unified analysis of random fourier features. *J. Mach. Learn. Res.*, 22(108):1–51, 2021.

F. Liu, X. Huang, Y. Chen, and J. A. K. Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. Preprint arXiv:2004.11154, 2020.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci.*, 115(33), 2018.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, pages 2388–2464, 2019.

S. Mendelson. Upper bounds on product and multiplier empirical processes. *Stoch. Proc. Appl.*, 126(12):3652–3680, 2016.

D. Needell, A. A. Nelson, R. Saab, and P. Salanevich. Random vector functional link networks for function approximation on manifolds. Preprint arXiv:2007.15776, 2020.

A. Nitanda, G. Chinot, and T. Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *Preprint arXiv:1905.09870*, 2019.

S. Oymak and B. Recht. Near-optimal bounds for binary embeddings of arbitrary sets. Preprint arXiv:1512.04433, 2015.

S. Oymak and M. Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4951–4960, 2019.

Y. Plan and R. Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete Comput. Geom.*, 51(2):438–461, 2014.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184, 2007.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, pages 1313–1320, 2008.

A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30*, pages 3218–3228, 2017.

A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng. On random weights and unsupervised feature learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML)*, pages 1089–1096, 2011.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Netw.*, 61:85–117, 2015.

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

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

Y. Sun, A. Gilbert, and A. Tewari. On the approximation properties of random relu features. Preprint arXiv:1810.04374, 2018.

M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*, volume 3 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer Berlin Heidelberg, 2014.

R. Vershynin. Estimation in high dimensions: A geometric perspective. In G. E. Pfander, editor, *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, pages 3–66. Birkhäuser Cham, 2015.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018.

R. Vershynin. Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM J. Math. Data Sci.*, 2(4):1004–1033, 2020.

C. Xu and L. Jacques. Quantized compressive sensing with RIP matrices: the benefit of dithering. *Inf. Inference*, 9(3):543–586, 2020.

G. Yehudai and O. Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems 32*, 2019.

C. Yun, S. Sra, and A. Jadbabaie. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In *Advances in Neural Information Processing Systems 32*, 2019.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021.

D. Zou and Q. Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems 32*, 2019.