Benefits of Additive Noise in Composing Classes with Bounded Capacity

Anonymous Author(s)
Affiliation
Address
email

Abstract

We observe that given two (compatible) classes of functions $\mathcal{F}$ and $\mathcal{H}$ with small capacity as measured by their uniform covering numbers, the capacity of the composition class $\mathcal{H} \circ \mathcal{F}$ can become prohibitively large or even unbounded. We then show that adding a small amount of Gaussian noise to the output of $\mathcal{F}$ before composing it with $\mathcal{H}$ can effectively control the capacity of $\mathcal{H} \circ \mathcal{F}$, offering a general recipe for modular design. To prove our results, we define new notions of uniform covering number of random functions with respect to the total variation and Wasserstein distances. We instantiate our results for the case of multi-layer neural networks. Preliminary empirical results indicate that the amount of noise required for our bound to improve over existing uniform bounds can be quite low.

1 Introduction

Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$, and $\mathcal{H}$ a class of functions from $\mathcal{Y}$ to $\mathcal{Z}$. Assuming that $\mathcal{F}$ and $\mathcal{H}$ have bounded "capacity", can we bound the capacity of their composition, i.e., $\mathcal{H} \circ \mathcal{F} = \{h \circ f \mid f \in \mathcal{F}, h \in \mathcal{H}\}$? Here, by capacity we mean learning-theoretic quantities such as VC dimension, fat-shattering dimension, and (uniform) covering numbers associated with these classes (see Vapnik (1999); Anthony et al. (1999); Shalev-Shwartz and Ben-David (2014); Mohri et al. (2018) for an introduction). Being able to control the capacity of composition of function classes is useful, as it offers a modular approach to design sophisticated classes (and therefore learning algorithms) out of simpler ones. To be concrete, we want to know if the uniform covering number (as defined in the next section) of $\mathcal{H} \circ \mathcal{F}$ can be "effectively" bounded as a function of the uniform covering numbers of $\mathcal{F}$ and $\mathcal{H}$.

The answer to the above questions is true when $\mathcal{F}$ is a set of binary valued functions (i.e., $\mathcal{Y} = \{0, 1\}$ in the above). More generally, the capacity of the composition class (as measured by the uniform covering number) can be bounded as long as $|\mathcal{Y}|$ is relatively small (see Proposition 7). But what if $\mathcal{Y}$ is an infinite set, such as the natural case of $\mathcal{Y} = [0, 1]$? Unfortunately, in this case the capacity of $\mathcal{H} \circ \mathcal{F}$ (as measured by the covering number) can become unbounded (or excessively large) even when both $\mathcal{F}$ and $\mathcal{H}$ have bounded (or small) capacities; see Propositions 8 and 9.

Given the above observation, we ask whether there is a general and systematic way to control the capacity of the composition of bounded-capacity classes. More specifically, we are interested in the case where the domain sets are multi-dimensional real-valued vectors (e.g., $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}^p$, $\mathcal{Z} \subset \mathbb{R}^q$). The canonical examples of such classes are those associated with neural networks.

A common approach to control the capacity of $\mathcal{H} \circ \mathcal{F}$ is assuming that $\mathcal{H}$ and $\mathcal{F}$ have bounded capacity and $\mathcal{H}$ consists of Lipschitz functions (with respect to appropriate metrics). Then the capacity of $\mathcal{H} \circ \mathcal{F}$ can be bounded as long as $\mathcal{H}$ has a small "global cover" (see Remark 13). This observation has been used to bound the capacity of neural networks in terms of the magnitude of

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

their weights (Bartlett, 1996). More generally, the capacity of neural networks that admit Lipschitz continuity can be bounded based on their group norms and spectral norms (Neyshabur et al., 2015; Bartlett et al., 2017; Golowich et al., 2018). One benefit of this approach is that the composition of Lipschitz classes is still Lipschitz (although with a larger Lipschitz constant).

While building classes of functions from composition of Lipschitz classes is useful, it does not necessarily work as a general recipe. In fact, some commonly used classes of functions do not admit a small Lipschitz constant. Consider the class of single-layer neural networks defined over bounded input domain $[-B, B]^d$ and with the sigmoid activation function. While the sigmoid activation function itself is Lipschitz, the Lipschitz constant of the network depends on the magnitude of the weights. Indeed, we empirically observe that this can turn Lipschitzness-based bounds on the covering number of neural networks worse than classic VC-based bounds.

Another limitation of using Lipschitz classes is that they cannot be easily "mixed and matched" with other (bounded-capacity) classes. For example, suppose $\mathcal{F}$ is a class of L-Lipschitz functions (e.g., multi-layer sigmoid neural networks with many weights but small magnitudes). Also, assume $\mathcal{H}$ is a non-Lipschitz class with bounded uniform covering number (e.g., one layer sigmoid neural network with unbounded weights). Then although both $\mathcal{F}$ and $\mathcal{H}$ have bounded capacity, $\mathcal{H} \circ \mathcal{F}$ is not Lipschitz and its capacity cannot be generally controlled.

We take a different approach for composing classes of functions. A key observation that we make and utilize is that adding a little bit of noise while "gluing" two classes can help in controlling the capacity of their composition. In order to prove such results, we define and study uniform covering numbers of random functions with respect to total variation and Wasserstein metrics. The bounds for composition then come naturally through the use of data processing inequality for the total variation distance metric.

Contributions and Organization.

- Section 3 provides the necessary notations and includes the observations that composing real-valued functions can be more challenging than binary valued functions (Propositions 7,8, and 9).
- In Section 4, we define a new notion of covering number for random functions (Definition 10) with respect to total variation (TV) and Wasserstein distances.
- The bulk of our technical results appear in Section 5. These include a composition result for random classes with respect to the TV distance (Lemma 18) that is based on the data processing inequality. We also show how one can translate TV covering numbers to conventional $\|.\|_2$ counterparts (Theorem 17) and vice versa (Corollary 21). A useful tool is Theorem 20 which exploits kernel density estimation techniques to translate Wasserstein covers to TV covers when we add Gaussian noise to the output of functions.
- Section 6 provides a stronger type of covering number for classes of single-layer noisy neural networks with the sigmoid activation function (Theorem 25).
- In Section 7, we use the tools developed in the previous sections and prove a novel bound on the $\|.\|_2$ covering number of noisy deep neural networks (Theorem 26).
- In Section 8 we define NVAC, a metric for comparing generalization bounds (Definition 28) based on the number of samples required to make the bound non-vacuous.
- We offer some preliminary experiments, comparing various generalization bounds in Section 9. We observe that even a negligible amount of Gaussian noise can improve NVAC over other approaches without affecting the accuracy of the model on train or test data.

2 Related work

Adding various types of noise have been empirically shown to be beneficial in training neural networks. In dropout noise (Srivastava et al., 2014) (and its variants such as DropConnect (Wan et al., 2013)) the output of some of the activation functions (or weights) are randomly set to zero. These approaches are thought to act as a regularizer. Another example is Denoising AutoEncoders (Vincent et al., 2008) which adds noise to the input of the network while training stacked autoencoders.

There has been efforts on studying the theory behind the effects of noise in neural networks. Jim et al. (1996) study the effects of different types of additive and multiplicative noise on convergence speed

and generalization of recurrent neural networks (RNN) and suggest that noise can help to speed up the convergence on local minima surfaces. Lim et al. (2021) formalize the regularization effects of noise in RNNs and show that noisy RNNs are more stable and robust to input perturbations. Wang et al. (2019) and Gao and Zhou (2016) analyze the networks with dropout noise and find bounds on Rademacher complexities that are dependent on the product of norms and dropout probability. It is noteworthy that our techniques and results are quite different, and require a negligible amount of additive noise to work, while existing bounds for dropout improve over conventional bounds only if the amount of noise is substantial. Studying dropout noise with the tools developed in this paper is a direction for future research.

Studying PAC learning and its sample complexity is by now a mature field; see Vapnik (1999); Shalev-Shwartz and Ben-David (2014); Mohri et al. (2018). In the case of neural networks, standard Vapnik-Chervonenkis-based complexity bounds have been established (Baum and Haussler, 1988; Maass, 1994; Goldberg and Jerrum, 1995; Vidyasagar, 1997; Sontag et al., 1998; Koiran and Sontag, 1998; Bartlett et al., 1998; Bartlett and Maass, 2003; Bartlett et al., 2019). These offer generalization bounds that depend on the number of parameters of the neural network. There is also another line of work that aims to prove a generalization bound that mainly depends on the norms of the weights and Lipschitz continuity properties of the network rather than the number of parameters (Bartlett, 1996; Anthony et al., 1999; Zhang, 2002; Bartlett, 1996; Neyshabur et al., 2015; Bartlett et al., 2017; Neyshabur et al., 2018; Golowich et al., 2018; Arora et al., 2018; Nagarajan and Kolter, 2018; Long and Sedghi, 2020). We provide a more detailed discussion of some of these results in Appendix H. Finally, we refer the reader to Anthony et al. (1999) for an introductory discussion on this subject.

The above-mentioned bounds are usually vacuous for commonly used data sets and architectures. Dziugaite and Roy (2017) (and later Zhou et al. (2019)) show how to achieve a non-vacuous bound using the PAC Bayesian framework. These approaches as well as compression-based methods (Arora et al., 2018) are, however, examples of "two-step" methods; see Appendix H for more details. It has been argued that uniform convergence theory may not fully explain the performance of neural networks (Zhang et al., 2021; Nagarajan and Kolter, 2019). One conjecture is that implicit bias of gradient descent (Gunasekar et al., 2017; Arora et al., 2019; Ji et al., 2020; Chizat and Bach, 2020; Ji and Telgarsky, 2021) can lead to benign overfitting (Belkin et al., 2018, 2019; Bartlett et al., 2020); see Bartlett et al. (2021) for a recent overview.

3 Notations and background

Notation. $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^p$ denote two (domain) sets. For $x \in \mathcal{X}$, let $\|x\|_1$, $\|x\|_2$, and $\|x\|_\infty$ denote the ℓ_1, ℓ_2, and ℓ_∞ norm of the vector x, respectively. We denote the cardinality of a set S by $|S|$. The set of natural numbers smaller or equal to m are denoted by $[m]$. A hypothesis is a Borel function $f : \mathbb{R}^d \to \mathbb{R}^p$, and a hypothesis class $\mathcal{F}$ is a set of hypotheses.

We also define the random counterparts of the above definitions and use an overline to distinguish them from the non-random versions. $\overline{\mathcal{X}}$ denotes the set of all absolutely continuous random variables defined over $\mathcal{X}$. We sometimes abuse the notation and write $\overline{x} \in \mathcal{X}$ rather than $\overline{x} \in \overline{\mathcal{X}}$ (e.g., $\overline{x} \in \mathbb{R}^d$ is a random variable taking values in $\mathbb{R}^d$). By $\overline{y} = f(\overline{x})$ we denote a random variable that is the result of mapping $\overline{x}$ using a Borel function $f : \mathbb{R}^d \to \mathbb{R}^p$. We use $\overline{f} : \mathbb{R}^d \to \mathbb{R}^p$ to indicate that the mapping itself can be random. We use $\overline{\mathcal{F}}$ to signal that the class can include random hypotheses. We conflate the notation for random hypotheses so that they can be applied to both random and non-random inputs (e.g., $\overline{f}(\overline{x})$ and $\overline{f}(x)$).[1]

Definition 1 (Composition of two hypothesis classes)**.** *We denote by $h \circ f$ the function $h(f(x))$ (assuming the range of f and the domain of h are compatible). The composition of two hypothesis classes $\mathcal{F}$ and $\mathcal{H}$ is defined by $\mathcal{H} \circ \mathcal{F} = \{h \circ f \mid h \in \mathcal{H}, f \in \mathcal{F}\}$. Composition of classes of random hypotheses is defined similarly by $\overline{\mathcal{H}} \circ \overline{\mathcal{F}} = \{\overline{h} \circ \overline{f} \mid \overline{h} \in \overline{\mathcal{H}}, \overline{f} \in \overline{\mathcal{F}}\}$.*

The following singleton class $\overline{\mathcal{G}_\sigma}$ will be used to create noisy functions (e.g., using $\overline{\mathcal{G}_\sigma} \circ \mathcal{F}$).

Definition 2 (The Gaussian Noise Class)**.** *The d-dimensional noise class with scale σ is denoted by $\overline{\mathcal{G}_{\sigma,d}} = \{\overline{g_{\sigma,d}}\}$. Here, $\overline{g_{\sigma,d}} : \mathbb{R}^d \to \mathbb{R}^d$ is a random function defined by $\overline{g_{\sigma,d}}(\overline{x}) = \overline{x} + \overline{z}$, where $\overline{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. When it is clear from the context we drop d and write $\overline{\mathcal{G}_\sigma} = \{\overline{g_\sigma}\}$.*

[1] Technically, we consider $\overline{f}(x)$ to be $\overline{f}(\overline{\delta_x})$, where $\overline{\delta_x}$ is a random variable with Dirac delta measure on x.

In the rest of this section, we define the standard notion of uniform covering numbers for hypothesis classes. Intuitively, classes with larger uniform covering numbers have more capacity/flexibility, and therefore require more samples to be learned.

Definition 3 (Covering number)**.** *Let $(\mathcal{X}, \rho)$ be a metric space. We say that a set $A \subset \mathcal{X}$ is ϵ-covered by a set $C \subseteq A$ with respect to ρ, if for all $a \in A$ there exists $c \in C$ such that $\rho(a, c) \leq \epsilon$. The cardinality of the smallest set C that ϵ-covers A is denoted by $N(\epsilon, A, \rho)$ and it is referred to as the ϵ-covering number of A with respect to metric ρ.*

Definition 4 (Extended metrics)**.** *Let $(\mathcal{X}, \rho)$ be a metric space. Let $u = (a_1, \ldots, a_m), v = (b_1, \ldots, b_m) \in \mathcal{X}^m$ for $m \in \mathbb{N}$. The ∞-extended and ℓ_2-extended metrics over $\mathcal{X}^m$ are defined by $\rho^{\infty,m}(u, v) = \sup_{1 \leq i \leq m} \rho(a_i, b_i)$ and $\rho^{\ell_2,m}(u, v) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\rho(a_i, b_i))^2}$, respectively. We drop m and use ρ^∞ or ρ^{ℓ_2} if it is clear from the context.*

Remark 5. *The extended metrics are used in Definition 6 and capture the distance of two hypotheses on an input sample of size m. A typical example of ρ is the Euclidean distance over $\mathbb{R}^p$, for which the extended metrics are denoted by $\|.\|_2^{\infty,m}$ and $\|.\|_2^{\ell_2,m}$. Unlike ∞-extended metric, the ℓ_2-extended metric is normalized by $1/\sqrt{m}$, and therefore we have $\rho^{\ell_2,m}(u, v) \leq \rho^{\infty,m}(u, v)$ for all $u, v \in \mathcal{X}^m$.*

Definition 6 (Uniform covering number)**.** *Let $(\mathcal{Y}, \rho)$ be a metric space and $\mathcal{F}$ a hypothesis class of functions from $\mathcal{X}$ to $\mathcal{Y}$. For a set of inputs $S = \{x_1, x_2, \ldots, x_m\} \subseteq \mathcal{X}$, we define the restriction of $\mathcal{F}$ to S as $\mathcal{F}_{|S} = \{(f(x_1), f(x_2), \ldots, f(x_m)) : f \in \mathcal{F}\} \subseteq \mathcal{Y}^m$. The uniform ϵ-covering numbers of hypothesis class $\mathcal{F}$ with respect to metrics $\rho^\infty, \rho^{\ell_2}$ are denoted by $N_U(\epsilon, \mathcal{F}, m, \rho^\infty)$ and $N_U(\epsilon, \mathcal{F}, m, \rho^{\ell_2})$ and are the maximum values of $N(\epsilon, \mathcal{F}_{|S}, \rho^{\infty,m})$ and $N(\epsilon, \mathcal{F}_{|S}, \rho^{\ell_2,m})$ over all $S \subseteq \mathcal{X}$ with $|S| = m$, respectively.*

It is well-known that the Rademacher complexity and therefore the generalization gap of a class can be bounded based on logarithm of the uniform covering number. For sake of brevity, we defer those results to Appendix F. Therefore, our main object of interest is bounding (logarithm of) the uniform covering number. The following propositions show that there is a stark difference between classes of functions with finite range versus continuous valued functions when it comes to bounding the uniform covering number of composite classes; the proofs can be found in Appendix B.

Proposition 7. *Let $\mathcal{Y}$ be a finite domain ($|\mathcal{Y}| = k$) and $\rho(y, \hat{y}) = 1\{y \neq \hat{y}\}$ be a metric over $\mathcal{Y}$. For any class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ and any class $\mathcal{H}$ of functions from $\mathcal{Y}$ to $\mathbb{R}^d$ we have $N_U(\epsilon, \mathcal{H} \circ \mathcal{F}, m, \|.\|_2^\infty) \leq N_1 . N_U(\epsilon, \mathcal{H}, mN_1, \|.\|_2^\infty)$ where $N_1 = N_U(0.5, \mathcal{F}, m, \rho^\infty)$.*

Proposition 8. *Let $\mathcal{F} = \{f_w(x) = wx \mid w \in (0, 1), x \in (0, 1)\}$ be a class of functions and $\mathcal{H} = \{h(y) = 1/y \mid y \in (0, 1)\}$ be a singleton class. Then, $N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2}) \leq \lceil 2/\epsilon^2 \rceil$ and $N_U(\epsilon, \mathcal{H}, m, \|.\|_2^{\ell_2}) = 1$, but $N_U(\epsilon, \mathcal{H} \circ \mathcal{F}, m, \|.\|_2^{\ell_2})$ is unbounded.*

Proposition 9. *For every $\epsilon > \epsilon' > 0$, there exist hypothesis classes $\mathcal{F}$ and $\mathcal{H}$ such that for every m we have $N_U(\epsilon', \mathcal{H}, m, \|.\|_2^\infty) \leq m + 1$ and $N_U(\epsilon', \mathcal{F}, m, \|.\|_2^\infty) = 1$, yet $N_U(\epsilon, \mathcal{H} \circ \mathcal{F}, m, \|.\|_2^\infty) \geq 2^m$.*

4 Covering random hypotheses

We want to establish the benefits of adding (a little bit of) noise when composing hypothesis classes. Therefore, we need to analyze classes of *random* hypotheses. One way to do this is to replace each hypothesis with its expectation, creating a deterministic version of the hypothesis class. Unfortunately, this approach misses the whole point of having noisy hypotheses (and their benefits in composition). Instead, we extend the definition of uniform covering numbers to classes of random hypotheses $\overline{\mathcal{F}}$. The following is basically the random counterpart of Definition 6.

Definition 10 (Uniform covering number for classes of random hypotheses)**.** *Let $(\overline{\mathcal{Y}}, \rho)$ be a metric space and $\overline{\mathcal{F}}$ a class of random hypotheses from $\overline{\mathcal{X}}$ to $\overline{\mathcal{Y}}$. For a set of random variables $\overline{S} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_m}\} \subseteq \overline{\mathcal{X}}$, we define the restriction of $\overline{\mathcal{F}}$ to $\overline{S}$ as $\overline{\mathcal{F}}_{|\overline{S}} = \{(\overline{f}(\overline{x_1}), \overline{f}(\overline{x_2}), \ldots, \overline{f}(\overline{x_m})) : \overline{f} \in \overline{\mathcal{F}}\} \subseteq \overline{\mathcal{Y}}^m$. Let $\Gamma \subseteq \overline{\mathcal{X}}$. The uniform ϵ-covering numbers of $\overline{\mathcal{F}}$ with respect to Γ and metrics ρ^∞ and ρ^{ℓ_2} are defined by*

$$N_U(\epsilon, \overline{\mathcal{F}}, m, \rho^\infty, \Gamma) = \sup_{S \subseteq \Gamma, |S| = m} N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, \rho^{\infty,m}),$$

$$N_U(\epsilon, \overline{\mathcal{F}}, m, \rho^{\ell_2}, \Gamma) = \sup_{S \subseteq \Gamma, |S| = m} N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, \rho^{\ell_2,m}).$$

Remark 11. *Unlike in Definition 6 where ρ is usually the $\|.\|_2$ metric in the Euclidean space, here in Definition 10 ρ is defined over random variables. More specifically, we will use the Total Variation and Wasserstein metrics as concrete choices for ρ.*

Remark 12. *The specific choices that we use for Γ are*

- $\Gamma = \overline{\mathcal{X}_d}$*: the set of all absolutely continuous random variables defined over* $\mathbb{R}^d$.
- $\Gamma = \overline{\mathcal{X}_{B,d}}$*: the set of all absolutely continuous random variables defined over* $[-B, B]^d$.
- $\Gamma = \overline{\Delta_d} = \{\overline{\delta_x} \mid x \in \mathbb{R}^d\}$ *and* $\Gamma = \overline{\Delta_{B,d}} = \{\overline{\delta_x} \mid x \in [-B, B]^d\}$*, where* $\overline{\delta_x}$ *is the random variable associated with Dirac delta measure on* x.
- $\Gamma = \overline{\mathcal{G}_{\sigma,d}} \circ \overline{\mathcal{X}_{B,d}} = \{\overline{g_{\sigma,d}}(\overline{x}) \mid \overline{x} \in \overline{\mathcal{X}_{B,d}}\}$*: all members of* $\overline{\mathcal{X}_{B,d}}$ *after being "smoothed" by adding (convolving with) Gaussian noise.*

Remark 13. *Some hypothesis classes that we work with have "global" covers, in the sense that the uniform covering number does not depend on m. We therefore use the following notation*

$$N_U(\epsilon, \overline{\mathcal{F}}, \infty, \rho^\infty, \Gamma) = \lim_{m \to \infty} N_U(\epsilon, \overline{\mathcal{F}}, m, \rho^\infty, \Gamma).$$

We now define Total Variation (TV) and Wasserstein metrics over probability measures rather than random variables, but with a slight abuse of notation we will use them for random variables too.

Definition 14 (Total Variation Distance)**.** *Let μ and ν denote two probability measures over $\mathcal{X}$ and let Ω be the Borel sigma-algebra over $\mathcal{X}$. The TV distance between μ and ν is defined by*

$$d_{TV}(\mu, \nu) = \sup_{B \in \Omega} |\mu(B) - \nu(B)|.$$

Furthermore, if μ and ν have densities f and g then

$$d_{TV}(\mu, \nu) = \sup_{B \in \Omega} \Big| \int_B (f(x) - g(x)) dx \Big| = \frac{1}{2} \int_{\mathcal{X}} |f(x) - g(x)| \, dx = \frac{1}{2} \|f - g\|_1.$$

Definition 15 (Wasserstein Distance)**.** *Let μ and ν denote two probability measures over $\mathcal{X}$, and $\Pi(\mu, \nu)$ be the set of all their couplings. The Wasserstein distance between μ and ν is defined by*

$$d_{\mathcal{W}}(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|_2 d\pi(x, y) \right).$$

The following proposition makes it explicit that the conventional uniform covering number with respect to $\|.\|_2$ (Definition 6) can be regarded as a special case of Definition 10.

Proposition 16. *Let $\mathcal{F}$ be a class of (deterministic) hypotheses from $\mathbb{R}^d$ to $\mathbb{R}^p$. Then*

$N_U(\epsilon, \mathcal{F}, m, \|.\|_2^\infty) = N_U(\epsilon, \mathcal{F}, d_{\mathcal{W}}^\infty, m, \overline{\Delta_d})$ *and* $N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2}) = N_U(\epsilon, \mathcal{F}, d_{\mathcal{W}}^{\ell_2}, m, \overline{\Delta_d})$.

The proposition is the direct consequence of the Definitions 6 and 10 once we note that the Wasserstein distance between Dirac random variables is just their ℓ_2 distance, i.e., $d_{\mathcal{W}}(\overline{\delta_x}, \overline{\delta_y}) = \|x - y\|_2$.

5 Bounding the uniform covering number

This section provides tools that can be used in a general recipe for bounding the uniform covering number. The ultimate goal is to bound the (conventional) $\|.\|_2^\infty$ and $\|.\|_2^{\ell_2}$ uniform covering numbers for (noisy) compositions of hypothesis classes. In order to achieve this, we will show how one can turn TV covers into $\|.\|_2$ covers (Theorem 17) and vice versa (Corollary 21). But what is the point of going back and forth between $\|.\|_2$ and TV covers? Basically, the data processing inequality ensures an effective composition (Lemma 18) for TV covers. Our analysis goes through a number of steps, connecting covering numbers with respect to $\|.\|_2$, Wasserstein, and TV distances. The missing proofs of this section can be found in Appendix C.

The following theorem considers the deterministic class $\mathcal{H}$ associated with expectations of random hypotheses from $\overline{\mathcal{F}}$, and shows that bounding the uniform covering number of $\mathcal{F}$ with respect to TV distance is enough for bounding the uniform covering number of $\mathcal{H}$ with respect to $\|.\|_2$ distance.

Theorem 17 (From a TV cover to a $\|.\|_2$ cover)**.** *Consider any class $\overline{\mathcal{F}}$ of random hypotheses $\overline{f} : \mathbb{R}^d \to [-B, B]^p$ with bounded output. Define the (nonrandom) hypothesis class $\mathcal{H} = \{h : \mathbb{R}^d \to [-B, B]^p \mid h(x) = \mathbb{E}_{\overline{f}}\left[\,\overline{f}(x)\right], \overline{f} \in \overline{\mathcal{F}}\}$. Then for every $\epsilon > 0$, $m \in \mathbb{N}$ these two inequalities hold:*

$$N_U(2B\epsilon\sqrt{p}, \mathcal{H}, m, \|.\|_2^\infty) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\Delta_d}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\mathcal{X}_d}),$$
$$N_U(2B\epsilon\sqrt{p}, \mathcal{H}, m, \|.\|_2^{\ell_2}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\mathcal{X}_d}).$$

But what is the point of working with the TV distance? An important ingredient of our analysis is the use of data processing inequality which holds for the TV distance (see Lemma 29). The following lemma uses this fact, and shows how one can compose classes with bounded TV covers.

Lemma 18 (Composing classes with bounded TV covers)**.** *Let $\overline{\mathcal{F}}$ be a class of random hypotheses from $\mathbb{R}^d$ to $\mathbb{R}^p$, and $\overline{\mathcal{H}}$ be a class of random hypotheses from $\mathbb{R}^p$ to $\mathbb{R}^q$. For every $\epsilon, \epsilon' > 0$, and every $m \in \mathbb{N}$ these three inequalities hold:*

$$N_U\left(\epsilon + \epsilon', \overline{\mathcal{H}} \circ \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\mathcal{X}_d}\right) \leq N_U\left(\epsilon', \overline{\mathcal{H}}, mN_1, d_{TV}^\infty, \overline{\mathcal{X}_p}\right).N_1,$$
$$N_U\left(\epsilon + \epsilon', \overline{\mathcal{H}} \circ \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\Delta_d}\right) \leq N_U\left(\epsilon', \overline{\mathcal{H}}, mN_2, d_{TV}^\infty, \overline{\mathcal{X}_p}\right).N_2,$$
$$N_U\left(\epsilon + \epsilon', \overline{\mathcal{H}} \circ \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}\right) \leq N_U\left(\epsilon', \overline{\mathcal{H}}, mN_3, d_{TV}^\infty, \overline{\mathcal{X}_p}\right).N_3,$$

where $N_1 = N_U\left(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\mathcal{X}_d}\right)$, $N_2 = N_U\left(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\Delta_d}\right)$ and $N_3 = N_U\left(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}\right)$.

Remark 19. *In Lemma 18, for $\overline{\mathcal{H}}$, we required the stronger notion of cover with respect to $\overline{\mathcal{X}_d}$ (i.e., the input to the hypotheses can be any random variable with a density function), whereas for $\overline{\mathcal{F}}$ a cover with respect to $\overline{\Delta_d}$ sufficed in some cases. As we will see below, finding a cover with respect to $\overline{\Delta_d}$ is easier since one can reuse conventional $\|.\|_2$ covers. However, finding covers with respect to $\overline{\mathcal{X}_d}$ is more challenging. In the next section we show how to do this for a class of neural networks.*

The next step is bounding the uniform covering number with respect to the TV distance (TV covering number for short). It will be useful to be able to bound TV covering number with Wasserstein covering number. However, this is generally impossible since closeness in Wasserstein distance does not imply closeness in TV distance. Yet, the following theorem establishes that one can bound the TV covering number as long as some Gaussian noise is added to the output of the hypotheses.

Theorem 20 (From a Wasserstein cover to a TV cover)**.** *Let $\overline{\mathcal{F}}$ be a class of random hypotheses from $\mathbb{R}^d$ to $\mathbb{R}^p$, and $\overline{\mathcal{G}_{\sigma,p}}$ be a Gaussian noise class. Then for every $\epsilon > 0$ and $m \in \mathbb{N}$ we have*

$$N_U\left(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_{\sigma,p}} \circ \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\mathcal{X}_d}\right) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{\mathcal{W}}^\infty, \overline{\mathcal{X}_d}),$$
$$N_U\left(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_{\sigma,p}} \circ \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\Delta_d}\right) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{\mathcal{W}}^\infty, \overline{\Delta_d}).$$

Intuitively, the Gaussian noise smooths out densities of random variables that are associated with applying transformation in $\overline{\mathcal{F}}$ to random variables in $\overline{\mathcal{X}_d}$ or $\overline{\Delta_d}$. As a result, the proof of Theorem 20 has a kernel density estimation step on the smooth densities. Finally, we can use Proposition 16 to relate the Wasserstein covering number with the $\|.\|_2$ covering number. The following corollary is the result of Proposition 16 and Theorem 20.

Corollary 21 (From a $\|.\|_2$ cover to a TV cover)**.** *Let $\mathcal{F}$ be a class of hypotheses $f : \mathbb{R}^d \to \mathbb{R}^p$ and $\overline{\mathcal{G}_{\sigma,p}}$ be a Gaussian noise class. Then for every $\epsilon > 0$ and $m \in \mathbb{N}$ we have*

$$N_U(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_{\sigma,p}} \circ \mathcal{F}, m, d_{TV}^\infty, \overline{\Delta_d}) \leq N_U(\epsilon, \mathcal{F}, m, \|.\|_2^\infty),$$
$$N_U(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_{\sigma,p}} \circ \mathcal{F}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) \leq N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2}).$$

The following theorem shows that we can get a stronger notion of TV cover with respect to $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}$ from a $\|.\|_2$ global cover, given that some Gaussian noise is added to the output of hypotheses. However, finding a small $\|.\|_2$ global cover is usually a challenging task. The proof involves finding a Wasserstein covering number and using Theorem 20 to obtain TV covering number.

Theorem 22 (From a global $\|.\|_2$ cover to a global TV cover)**.** *Let $\mathcal{F}$ be a class of hypotheses $f : \mathbb{R}^d \to \mathbb{R}^p$ and $\overline{\mathcal{G}_{\sigma,p}}$ be a Gaussian noise class. Then for every $\epsilon > 0$ and $m \in \mathbb{N}$ we have*

$$N_U\left(\frac{9\epsilon}{\sigma}, \overline{\mathcal{G}_{\sigma,p}} \circ \mathcal{F}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_{\sigma,d}} \circ \overline{\mathcal{X}_{B,d}}\right) \leq N_U(\epsilon, \mathcal{F}, \infty, \|.\|_2^{\infty}).$$

6 Uniform TV covers for single-layer neural networks

In this section, we study the uniform covering number of single-layer neural networks with respect to the total variation distance. This will set the stage for the next section, where we want to use the tools from Section 5 to bound covering numbers of deeper networks. We start with the following definition for the class of single-layer neural networks.

Definition 23 (Single-Layer Sigmoid Neural Networks)**.** *Let $\Phi : \mathbb{R}^p \to [0,1]^p$ be the element-wise sigmoid activation function defined by $\Phi((x^{(1)}, \ldots, x^{(p)})) = (\phi(x^{(1)}), \ldots, \phi(x^{(p)}))$, where $\phi(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. The class of single-layer neural networks with d inputs and p outputs is defined by NET[d, p] $= \{f_W : \mathbb{R}^d \to [0,1]^p \mid f_W(x) = \Phi(W^\top x), W \in \mathbb{R}^{d \times p}\}$.*

Remark 24. *We choose sigmoid function for simplicity, but our analysis for finding uniform covering numbers of neural networks (Theorem 25) is not specific to the sigmoid activation function. We present a stronger version of Theorem 25 in Appendix D which works for any activation function that is Lipschitz, monotone, and bounded.*

As mentioned in Remark 19, Lemma 18 requires stronger notion of covering numbers with respect to $\overline{\mathcal{X}_d}$ and TV distance. In fact, the size of this kind of cover is infinite for deterministic neural networks defined above. In contrast, Theorem 25 shows that one can bound this covering number as long as some Gaussian noise is added to the input and output of the network. The proof is quite technical, starting with estimating the smoothed input distribution ($\overline{g_\sigma}(x)$) with mixtures of Gaussians using kernel density estimation. Then a cover for mixtures of Gaussians with respect to Wasserstein distance is found. Finally, Theorem 20 helps to find the cover with respect to total variation distance. For a complete proof of theorem see Appendix D.

Theorem 25 (A global total variation cover for noisy neural networks with unbounded weights)**.** *For every $p, d \in \mathbb{N}, \epsilon > 0, \sigma < 30d/\epsilon$ we have*

$$N_U(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{NET}[d,p], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{1,d}}) \leq \left(282 \frac{d^{5/2}\sqrt{\ln\left((30d - \epsilon\sigma)/(\epsilon\sigma)\right)}}{\epsilon^{3/2}\sigma^2} \ln\left(\frac{30d}{\epsilon\sigma}\right)\right)^{p(d+1)}.$$

Note that the dependence of the bound on $1/\sigma$ is polynomial. The assumption $\sigma \ll 30d/\epsilon$ holds for any reasonable application (we will use $\sigma \ll 1$ in the experiments). In contrast to the analyses that exploit Lipschitz continuity, the above theorem does not require any assumptions on the norms of weights. Theorem 25 is a key tool in analyzing the uniform covering number of deeper networks.

7 Uniform covering numbers for deeper networks

In the following, we discuss how one can use Theorem 25 and techniques provided in Section 5 to obtain bounds on covering number for deeper networks. For a T-layer neural network, it is useful to separate the first layer from the rest of the network. The following theorem offers a bound on the uniform covering number of (the expectation of) a noisy network based on the usual $\|.\|_2^{\ell_2}$ covering number of the first layer and the TV covering number of the subsequent layers.

Theorem 26. *Let NET[d, p_1], NET[p_1, p_2], $\ldots$, NET[p_{T-1}, p_T] be T classes of neural networks. Denote the T-layer noisy network by*

$$\overline{\mathcal{F}} = \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{T-1}, p_T] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_1, p_2] \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[d, p_1],$$

and let $\mathcal{H} = \{h : \mathbb{R}^d \to [0,1]^{p_T} \mid h(x) = \mathbb{E}_{\overline{f}}\left[\overline{f}(x)\right], \overline{f} \in \overline{\mathcal{F}}\}$. Denote the uniform covering numbers of compositions of neural network classes with the Gaussian noise class (with respect to d_{TV}^{∞}) as

$$N_i = N_U\left(\frac{\epsilon}{2T\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{i-1}, p_i], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{1,p_{i-1}}}\right), \ 2 \leq i \leq T, \tag{1}$$

and the uniform covering number of $\overline{\mathcal{G}_\sigma} \circ$ *NET*$[d, p_1]$ *with respect to* $\|.\|_2^{\ell_2}$ *as*

$$N_1 = N_U\left(\frac{2\sigma\epsilon}{18T\sqrt{p_T}}, \textit{NET}[d, p_1], m, \|.\|_2^{\ell_2}\right).$$

Then we have

$$N_U\left(\epsilon, \mathcal{H}, m, \|.\|_2^{\ell_2}\right) \leq \prod_{i=1}^{T} N_i.$$

The $\|.\|_2^{\ell_2}$ covering number of the first layer (i.e., N_1 in above) can be bounded using standard approaches in the literature. For instance, in Appendix H we will use the bound of Lemma 14.7 in Anthony et al. (1999). Other N_i's can be bounded using Theorem 25. The above bound does not depend on the norm of weights and therefore we can use it for networks with large weights.

The proof of Theorem 26 involves applying Corollary 21 to turn the $\|.\|_2$ cover of first layer into a TV cover. We then find a TV cover for rest of the network by combining Theorem 26 and Lemma 18. We will compose the first layer with the rest of the network and bound the covering number by another application of Lemma 18. Finally, we turn the TV covering number (of the entire network) back into $\|.\|_2^{\ell_2}$ covering number using Theorem 17. The complete proof can be found in Appendix E.

One can generalize the above analysis in the following way: instead of separating the first layer, one can basically "break" the network from any layer, use existing $\|.\|_2$ covering number bounds for the first few layers, and Theorem 25 for the rest. See Lemma 35 in Appendix E for details.

8 NVAC: a metric for comparing generalization bounds

We want to provide tools to compare different approaches in finding covering numbers and their suggested generalization bounds. First, we define the notion of a generalization bound for classification. Let $\mathcal{Y} = [k]$ and $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}^k$. Let $\mathcal{A}$ be an algorithm that receives a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ and outputs a function $\hat{h} \in \mathcal{F}$. Note that the output of this function is a real vector so it can capture margin-based classifiers too. Let $l^{0-1} : \mathbb{R}^k \times [k] \to \{0, 1\}$ be the "thresholded" 0-1 loss function defined by $l^{0-1}(u, y) = 1\{\text{argmax}_i u^{(i)} \neq y\}$ where $u^{(i)}$ is the i-th dimension of u.

Definition 27 (Generalization Bound for Classification)**.** *A (valid) generalization bound for* $\mathcal{A}$ *with respect to* l^{0-1} *and another (surrogate) loss function* l *is a function* $GB : \mathcal{F} \times (\mathcal{X} \times \mathcal{Y})^m \to \mathbb{R}$ *such that for every distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$*, if* $S \sim \mathcal{D}^m$*, then with probability at least 0.99 (over the randomness of* S*) we have*

$$\left|\frac{1}{m}\sum_{(x,y)\in S} l(\hat{h}(x), y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l^{0-1}(\hat{h}(x), y)\right]\right| \leq GB(\hat{h}, S).$$

For example, $GB(\hat{h}, S) = 2$ is a useless but valid generalization bound. Various generalization bounds that have been proposed in the literature are examples of a GB. Note that GB can depend both on S (for instance on $|S|$) and on $\hat{h}$ (for example, on the norm of the weights of network).

It is not straightforward to empirically compare generalization bounds since they are often vacuous for commonly used applications. Jiang et al. (2019) address this by looking at other metrics, such as the correlation of each bound with the actual generalization gap. While these metrics are informative, it is also useful to know how far off each bound is from producing a "non-vacuous" bound (Dziugaite and Roy, 2017). Therefore, we will take a more direct approach and propose the following metric.

Definition 28 (NVAC)**.** *Let* $\hat{h}$ *be a hypothesis,* $S \in (\mathcal{X} \times \mathcal{Y})^m$ *a sample, and* GB *a generalization bound for algorithm* $\mathcal{A}$*. Let* S^n *denote a sample of size* mn *which includes* n *copies of* S*. Let* n^* *be the smallest integer such that the following holds:*

$$GB(\hat{h}, S^{n^*}) + \frac{1}{|S^{n^*}|}\sum_{(x,y)\in S^{n^*}} l(\hat{h}(x), y) \leq 1.$$

We define NVAC to be $|S^{n^*}| = mn^*$.

Informally speaking, NVAC is an upper bound on the minimum number of samples required to obtain a non-vacuous generalization bound. Approaches that get tighter upper bounds on covering number will generally result in smaller NVACs. In Appendix G, we will show how one can calculate NVAC using the uniform covering number bounds.

9 Experiments

In this section, we empirically compare different approaches in bounding the covering number using the NVAC metric. We compare the following approaches in bounding covering number: Theorem 26, Norm-based (Theorem 14.17 in Anthony et al. (1999)), Lipschitzness-based (Theorem 14.5 in Anthony et al. (1999)), Pseudo-dim-based (Theorem 14.2 in Anthony et al. (1999)), and Spectral (Bartlett et al. (2017)). More details about these bounds can be found in Appendix H.

We train fully connected neural networks on MNIST dataset. We use a network with an input layer, an output layer, and three hidden layers each containing 250 hidden neurons as the baseline architecture. See Appendix I for the details of the learning settings. The left two graphs in Figure 1 depict NVACs as functions of the depth and width of the network. It can be observed that our approach achieves the smallest NVAC. The Norm-based bound is the worst and is removed from the graph (see Appendix I). Overall, bounds that are based on the norm of the weights (even the spectral norm) perform poorly compared to those that are based on the parameter count. This is an interesting observation since we have millions of parameters ($\approx 3 \times 10^9$) in some of the wide networks and one would assume approaches based on norm of weights should be able to explain generalization behaviour better. There are several reasons why our bound performs better. First, the dependence of NVAC on $1/\epsilon$ is linear for the Spectral approach and polynomial for Norm-based approach while the dependence is logarithmic in our approach. Second, norm-based bounds depend on product of norms and group norms which can get quite large. Finally, our method works naturally for multi-output layers, while the Pseudo-dim-based approach works for real-valued output (and therefore one needs to bound the cover for each output separately).

The covering number bound of Theorem 26 has a polynomial dependence on $1/\sigma$. Therefore, NVAC has a mild logarithmic dependence on $1/\sigma$ (see Appendix G for details). The third graph in Figure 1 corroborates that even a negligible amount of noise ($\sigma \approx 10^{-240}$) is sufficient to get tighter bounds on NVAC compared to other approaches. Finally, the right graph in Figure 1 shows that even with a considerable amount of noise (e.g, $\sigma = 0.2$), the train and test accuracy of the model remain almost unchanged. This is perhaps expected, as the dynamics of training neural networks with gradient descent is already noisy even without adding Gaussian noise. Overall, our preliminary experiment shows that small amount of noise does not affect the performance, yet it enables us to prove tighter generalization bounds.

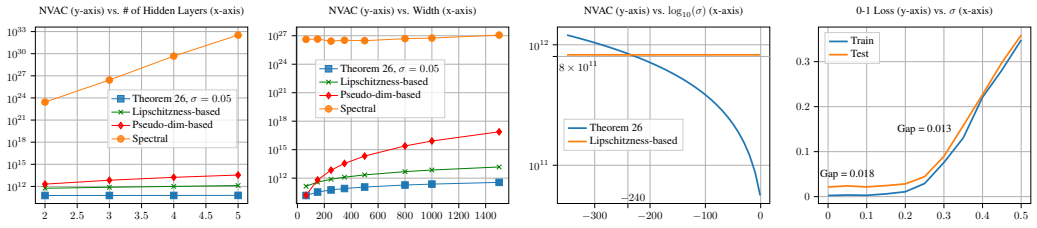

Figure 1: The left two graphs depict NVAC of different generalization bounds as a function of the number of hidden layers and width of the network. The Norm-based approach is excluded because of its excessively high NVAC (see Appendix I). The third graph plots NVAC against $\log_{10}(\sigma)$ (σ is standard deviation of noise) for the two best approaches. The rightmost graph plots the train/test 0-1 losses for different values of σ. The gaps between the train and test losses are shown for $\sigma = 0, 0.3$.

Limitations and Future Work. Our analysis is based on the assumption that the activation function is bounded. Therefore, extending the results to ReLU neural networks is not immediate, and is left for future work. Also, our empirical analysis is preliminary and is mostly used as a sanity check. Further empirical evaluations can help to better understand the role of noise in training neural networks.

References

Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9, 1996.

Peter Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc dimension bounds for piecewise polynomial networks. *Advances in neural information processing systems*, 11, 1998.

Peter L Bartlett and Wolfgang Maass. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, pages 1188–1192, 2003.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Eric Baum and David Haussler. What size net gives valid generalization? *Advances in neural information processing systems*, 1, 1988.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Minwoo Chae and Stephen G Walker. Wasserstein upper bounds of the total variation for smooth densities. *Statistics & Probability Letters*, 163:108771, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Richard M Dudley. Universal donsker classes and metric entropy. In *Selected Works of RM Dudley*, pages 345–365. Springer, 2010.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Wei Gao and Zhi-Hua Zhou. Dropout rademacher complexity of deep neural networks. *Science China Information Sciences*, 59(7):1–12, 2016.

Paul W Goldberg and Mark R Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2):131–148, 1995.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Kam-Chuen Jim, C Lee Giles, and Bill G Horne. An analysis of noise in recurrent neural networks: convergence and generalization. *IEEE Transactions on neural networks*, 7(6):1424–1438, 1996.

Pascal Koiran and Eduardo D Sontag. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Soon Hoe Lim, N Benjamin Erichson, Liam Hodgkinson, and Michael W Mahoney. Noisy recurrent neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Philip M Long and Hanie Sedghi. Size-free generalization bounds for convolutional neural networks. In *International Conference on Learning Representations*, 2020.

Wolfgang Maass. Neural nets with superlinear vc-dimension. *Neural Computation*, 6(5):877–884, 1994.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Vaishnavh Nagarajan and Zico Kolter. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Eduardo D Sontag et al. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Mathukumalli Vidyasagar. *A theory of learning and generalization: with applications to neural networks and control systems*. Springer-Verlag, 1997.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

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

Haotian Wang, Wenjing Yang, Zhenyu Zhao, Tingjin Luo, Ji Wang, and Yuhua Tang. Rademacher dropout: An adaptive dropout for deep neural network via optimizing generalization gap. *Neurocomputing*, 357:177–187, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *International Conference on Learning Representations (ICLR)*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The main claims of the paper in abstract and introduction are stated as theorems, lemmas and propositions, which are proved in appendices.
   (b) Did you describe the limitations of your work? [Yes] We discuss the limitations of our experiments and possible future work in "Limitations and Future Work".
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a mainly theoretical paper and is unlikely to have any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The theorems and lemmas are proved in appendices.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The codes will be provided as a supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix I.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We provide accuracy and error measures in the main paper and supplemental material.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix I

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We experiment on the publicly available MNIST dataset, which is mentioned in the main paper.

(b) Did you mention the license of the assets? [Yes] We provide our code as a supplemental material which does not require a license. We use MNIST dataset and discuss it in Appendix I.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provide our code in the supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use MNIST dataset which is a publicly available.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We use MNIST dataset which is a collection of handwritten digits and does not contain personal information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Miscellaneous facts

Lemma 29 (Data processing inequality for TV distance)**.** *Given two random variables $\overline{x_1}, \overline{x_2} \in \overline{\mathcal{X}}$, and a (random) Borel function $f : \mathcal{X} \to \mathcal{Y}$,*

$$d_{TV}(f(\overline{x_1}), f(\overline{x_2})) \leq d_{TV}(\overline{x_1}, \overline{x_2}).$$

The next theorem, bounds the total variation distance between two Gaussian random variables.

Theorem 30 (Total variation distance between Gaussians with same covariance)**.** *Let $\mathcal{N}(\mu_1, \sigma^2 I_d)$ and $\mathcal{N}(\mu_2, \sigma^2 I_d)$ be two Gaussian random variables, where I_d is the d-by-d identity matrix. Then we have,*

$$d_{TV}(\mathcal{N}(\mu_1, \sigma^2 I_d), \mathcal{N}(\mu_2, \sigma^2 I_d)) \leq \frac{1}{2\sigma} \left\|\mu_1 - \mu_2\right\|_2 .$$

Proof. Form Pinsker's inequality we know that for any two distributions P and Q we have

$$d_{TV}(P, Q) \leq \sqrt{\frac{1}{2} d_{KL}(P, Q)}, \tag{2}$$

where $d_{KL}(P, Q)$ is the Kullback-Liebler (KL) divergence between P and Q. We can the find the KL divergence between $\mathcal{N}(\mu_1, \sigma^2 I_d)$ and $\mathcal{N}(\mu_2, \sigma^2 I_d)$ as (see e.g., Diakonikolas et al. (2019))

$$d_{KL}\left(\mathcal{N}(\mu_1, \sigma^2 I_d), \mathcal{N}(\mu_2, \sigma^2 I_d)\right) \leq \frac{1}{2\sigma^2} \|\mu_1 - \mu_2\|_2^2. \tag{3}$$

Combining Equations 2 and 3 concludes the result. □

Lemma 31. *Let $Y \sim \chi^2_n$ be a chi-squared random variable with n degrees of freedom. Then we have (Laurent and Massart, 2000)*

$$\mathbb{P}[Y - n \geq 2\sqrt{nt} + 2t] \leq e^{-t}.$$

Lemma 32. *Let $x = \sum_{i=1}^{m} w_i g_i$ be a random variable, where g_i are d-dimensional Gaussian random variables with means $\mu_i \in [-B, B]^d$ and covariance matrices of $\sigma^2 I_d$. We have*

$$\mathbb{P}\left[\|x\|_2 \geq (B+\sigma)\sqrt{d} + \sigma\sqrt{2t}\right] \leq e^{-t}.$$

Proof. We know that for any $R \in \mathbb{R}$

$$\mathbb{P}\left[\|x\|_2^2 \geq R^2\right] = \sum_{i=1}^{m} w_i \mathbb{P}\left[\|g_i\|_2^2 \geq R^2\right] = \sum_{i=1}^{m} w_i \mathbb{P}\left[\|\sigma y_i + \mu_i\|_2^2 \geq R^2\right] = \sum_{i=1}^{m} w_i \mathbb{P}\left[\|\sigma y_i + \mu_i\|_2 \geq R\right],$$

where $y_i \sim \mathcal{N}(0, I_d)$ are standard normal random variables. Using triangle inequality we can rewrite the above equation as

$$\mathbb{P}\left[\|x\|_2^2 \geq R^2\right] \leq \sum_{i=1}^{m} w_i \mathbb{P}\left[\|\sigma y_i\|_2 + \|\mu_i\|_2 \geq R\right] \leq \sum_{i=1}^{m} w_i \mathbb{P}\left[\|\sigma y_i\|_2 + B\sqrt{d} \geq R\right].$$

We can, therefore, conclude that

$$\mathbb{P}\left[\|x\|_2^2 \geq R^2\right] \leq \mathbb{P}\left[\|y_i\|_2^2 \geq \left(\frac{R - B\sqrt{d}}{\sigma}\right)^2\right].$$

Setting $R = (B+\sigma)\sqrt{d} + \sigma\sqrt{2t}$, we can write

$$\begin{aligned}
&\mathbb{P}\left[\|x\|_2 \geq (B+\sigma)\sqrt{d} + \sigma\sqrt{2t}\right] \\
&= \mathbb{P}\left[\|x\|_2^2 \geq \left((B+\sigma)\sqrt{d} + \sigma\sqrt{2t}\right)^2\right] \\
&\leq \mathbb{P}\left[\|y_i\|_2^2 \geq (\sqrt{d} + \sqrt{2t})^2\right] \\
&\leq \mathbb{P}\left[\|y_i\|_2^2 \geq d + 2t + 2\sqrt{dt}\right] \\
&\leq e^{-t}.
\end{aligned}$$

□

B Proofs of propositions in Section 3

B.1 Proof of Proposition 7

Proof. Fix an input set $S = \{x_1, \ldots, x_m\}$. Let $C = \{\hat{f}_{i|S} \mid \hat{f}_i \in \mathcal{F}, i \in [r_1]\}$ be 0.5-cover for $\mathcal{F}_{|S}$ with respect to ρ^∞. Therefore, given any $f_{|S} \in \mathcal{F}_{|S}$ there exists $\hat{f}_{i|S} \in C$ such that

$$\rho^\infty\left((f(x_1), \ldots, f(x_m)), (\hat{f}_i(x_1), \ldots, \hat{f}_i(x_m))\right) \leq 0.5 \tag{4}$$

Since $\rho\left(f(x), \hat{f}_i(x)\right) = 1\{f(x) \neq \hat{f}_i(x)\}$, Equation 4 suggests that $f(x_l) = \hat{f}_i(x_k)$ for any $k \in [m]$. Let $S' = \{\hat{f}_i(x_k) \mid i \in [r_1],\, k \in [m]\}$ and $C' = \{\hat{h_j}_{|S'} \mid \hat{h_j} \in \mathcal{H}, j \in [r_2]\}$ be an ϵ-cover for $\mathcal{H}_{|S'}$ with respect to $\|.\|_2^\infty$. We know that $|S'| \leq mr_1$. Denote $\hat{\mathcal{Q}} = \{\hat{h_j} \circ \hat{f}_i \mid i \in [r_1],\, j \in [r_2]\}$. We will prove that $\hat{\mathcal{Q}}_{|S}$ is an ϵ-cover for $(\mathcal{H} \circ \mathcal{F})_{|S}$ with respect to $\|.\|_2^\infty$. Consider $(h \circ f)_{|S} = (h(f(x_1)), \ldots, h(f(x_m))) \in (\mathcal{H} \circ \mathcal{F})_{|S}$. Since C is a 0.5-cover for $\mathcal{F}_{|S}$, from equation 4, we know that there exists $\hat{f}_i \in \mathcal{F}$ such that $f(x_k) = \hat{f}_i(x_k)$ for any $k \in [m]$. On the other hand, for any $k \in [m]$, $\hat{f}_i(x_k)$ is an element of S', consequently, there exists $\hat{h_j} \in \mathcal{H}$ such that

$$\begin{aligned}
&\left\|\left(h(\hat{f}_i(x_1)), \ldots, h(\hat{f}_i(x_m))\right) - \left(\hat{h_j}(\hat{f}_i(x_1)), \ldots, \hat{h_j}(\hat{f}_i(x_m))\right)\right\|_2^\infty \\
&= \left\|\left(h(f(x_1)), \ldots, h(f(x_m))\right) - \left(\hat{h_j}(\hat{f}_i(x_1)), \ldots, \hat{h_j}(\hat{f}_i(x_m))\right)\right\|_2^\infty \\
&\leq \epsilon
\end{aligned}$$

From the above equation, we can conclude that $(\mathcal{H} \circ \mathcal{F})_{|S}$ is ϵ-covered by $\hat{\mathcal{Q}}_{|S}$. Clearly, $\left|\hat{\mathcal{Q}}_{|S}\right| \leq r_1 r_2$ and we know that $mr_1 \leq mN_1$. Therefore, $N(\epsilon, \mathcal{H}_{|S'}, \|.\|_2^\infty) \leq N_U(\epsilon, \mathcal{H}, mr_1, \|.\|_2^\infty) \leq N_U(\epsilon, \mathcal{H}, mN_1, \|.\|_2^\infty)$ This result holds for any input set $S \subset \mathcal{X}^m$ with $|S| = m$, therefore, it follows that

$$N_U(\epsilon, \mathcal{H} \circ \mathcal{F}, m, \|.\|_2^\infty) \leq N_1 . N_U(\epsilon, \mathcal{H}, mN_1, \|.\|_2^\infty).$$

□

B.2 Proof of Proposition 8

Proof. The proof for the bound of $N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2})$ can be found under Theorem 3 in Zhang (2002). Since $\mathcal{H}$ is a singleton class, it is easy to verify $N_U(\epsilon, \mathcal{H}, m, \|.\|_2^\infty) = 1$. We prove that the covering number of $\mathcal{H} \circ \mathcal{F}$ is unbounded by contradiction. Let $S = \{x_1, \ldots, x_m\} \in (0,1)^m$ be an input set where $0 < x_1 \leq \ldots \leq x_m$. Denote $C = \{(h \circ \hat{f}_i)_{|S} = (\frac{1}{\hat{w}_i x_1}, \ldots, \frac{1}{w_i \hat{x}_m}) \mid \hat{f}_i \in \mathcal{F}, i \in [r]\}$ to be an ϵ-cover for $(\mathcal{H} \circ \mathcal{F})_{|S}$ where $|C| = r_1$ is finite. We know that $\hat{w}_i > 0$ for $i \in [r]$. Denote $w^* = \min_{i \in [r], \hat{w}_i > 0} \hat{w}_i$. Take any $w < \dfrac{1}{\frac{1}{w^*} + x_1 \epsilon} \leq \dfrac{1}{\frac{1}{\hat{w}_i} + x_1 \epsilon}$ and denote the corresponding function by $f \in \mathcal{F}$, i.e., $f(x) = wx$. we know that for every $i \in [r]$

$$\frac{1}{w x_1} > \frac{w}{\hat{w}_i x_1} + \epsilon.$$

This means that

$$\left\| \left(\frac{w}{x_1}, \ldots, \frac{w}{x_m} \right) - \left(\frac{\hat{w}_i}{x_1}, \ldots, \frac{\hat{w}_i}{x_m} \right) \right\|_2 = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\frac{w}{x_1} - \frac{\hat{w}_i}{x_1} \right)^2} \geq \epsilon$$

Therefore, there is no $(h \circ \hat{f}_i)_{|S} \in C$ such that $\|(h \circ \hat{f}_i)_{|S} - (h \circ f)_{|S}\|_2^{\ell_2} \leq \epsilon$, which contradicts with the assumption that C is an ϵ-cover for $(\mathcal{H} \circ \mathcal{F})_{|S}$. □

B.3 Proof of Proposition 9

Proof. Let $\mathcal{F}_{\gamma,\epsilon}$ denote the class of all functions $f_{\gamma,\epsilon}$ from $\mathcal{X}$ to $\mathbb{R}$ such that $|f(x) - x| \leq \gamma$ for any $x \in \mathcal{X}$, where $\gamma \leq \epsilon/2$. Fix an input set $S = \{x_1, \ldots, x_m\}$. We know that given any $f_{\gamma,\epsilon}, f'_{\gamma,\epsilon} \in \mathcal{F}_{\gamma,\epsilon}$ and $i \in [m]$,

$$\|f_{\gamma,\epsilon}(x_i) - f'_{\gamma,\epsilon}(x_i)\| \leq \|f_{\gamma,\epsilon}(x_i) - x_i\| + \|x_i - f'_{\gamma,\epsilon}(x_i)\| \leq \epsilon.$$

Therefore, it is easy to conclude that $N_U(\epsilon, \mathcal{F}_{\gamma,\epsilon}, m, \|.\|_2^\infty) = 1$. Let $\mathcal{H}$ to be the class of all threshold functions h_a from $\mathbb{R}$ to $[0,1]$, where $h_a(x) = 1\{x \geq a\}$. Consider an input set $S = \{x_1, \ldots, x_m\}$ where $x_1 \leq \ldots \leq x_m$. Given any $k \in [m]$ we can find $a \in \mathbb{R}$ such that $x_i < a$ for $1 \leq i \leq k$ and $x_i \geq a$ for $k < i \leq m$, e.g., set $a = (x_k + x_{k+1})/2$. We also know that for any $i, j \in [m]$, $h_a(x_i) \neq h_a(x_j)$ only if $x_i < a \leq x_j$. Therefore, it is easy to verify that $\mathcal{H}_{|S} = m + 1$ and that for any $h_{a|S}$ and $h_{a'|S}$ in $\mathcal{H}_{|S}$ we have $\left\|h_{a'|S} - h_{a|S}\right\|_2 \geq 1$. We can therefore conclude that $N_U(\epsilon, \mathcal{H}, m, \|.\|_2^\infty) = m + 1$. Next, consider the class $\mathcal{H} \circ \mathcal{F}_{\gamma,\epsilon}$. We prove that $N_U(\epsilon', \mathcal{H} \circ \mathcal{F}_{\gamma,\epsilon}, m, \|.\|_2^\infty) = 2^m$.

We first mention the fact that given any $(y_1, \ldots, y_m)$ and $(y'_1, \ldots, y'_m)$ in $\{0,1\}^m$ if there exists $i \in [m]$ such that $y_i \neq y'_i$, then $\|(y'_1, \ldots, y'_m) - (y_1, \ldots, y_m)\|_2 \geq 1$. Also, the range of the functions in $\mathcal{H} \circ \mathcal{F}$ is $[0,1]$, therefore, we are only interested in $\epsilon' < 1$. In the following, we prove that for any m there exists a set S' with $|S'| = m$ such that the restriction of $\mathcal{H} \circ \mathcal{F}$ to set S' has 2^m elements and the result follows.

Consider the input set $S' = \{z_1, \ldots, z_m\}$ such that $0 \leq z_1 < \ldots < z_m \leq \epsilon/2$. Given any $(y_1, \ldots, y_m) \in \{0,1\}^m$ we map $(z_1, \ldots, z_m)$ to $(e_1, \ldots, e_m)$ as follows: for any $i \in [m]$ if $y_i = 1$ we define $e_i = z_i + \epsilon/2$, otherwise we define $e_i = z_i - \epsilon/2$. This mapping can be done by some function $f_{\gamma,\epsilon}$ from $\mathcal{F}_{\gamma,\epsilon}$ since for any $i \in [m]$ we have $|e_i - z_i| = \epsilon/2$. Let $a = \epsilon/4$. We know that $h_a(e_i)$ is 1 if $y_i = 1$ and 0 otherwise. Therefore, we can conclude that for every element $(y_1, \ldots, y_m)$

in $\{0,1\}^m$, there exists $(h_a \circ f_{\gamma,\epsilon})_{|S'}$ in $(\mathcal{H} \circ \mathcal{F}_{\gamma,\epsilon})_{|S'}$ such that $(\mathcal{H} \circ \mathcal{F}_{\gamma,\epsilon})_{|S'} = (y_1, \ldots, y_m)$. Since $|\{0,1\}^m| = 2^m$, we can say that $N(\epsilon', (\mathcal{H} \circ \mathcal{F}_{\gamma,\epsilon})_{|S}, \|.\|_2^{\infty,m}) = 2^m$. Therefore,

$$N_U(\epsilon', \mathcal{H} \circ \mathcal{F}_{\gamma,\epsilon}, m, \|.\|_2^{\infty}) = \sup_{|S|=m} \left\{N(\epsilon', (\mathcal{H} \circ \mathcal{F}_{\gamma,\epsilon})_{|S}), \|.\|_2^{\infty,m})\right\} \geq 2^m.$$

□

C Proofs of theorems and lemmas in Section 5

Notation. For a (random) function f and an input set $S = \{x_1, \ldots, x_m\}$, we define the restriction of f to S as $f_{|S} = (f(x_1), \ldots, f(x_m))$. Therefore, the restriction of the class $\mathcal{F}$ to S can be denoted as $\mathcal{F}_{|S} = \{f_{|S} : f \in \mathcal{F}\}$. We also denote by $\mathscr{D}(\overline{x})$ the probability density functions of the random variable $\overline{x}$. For two Borel functions f_1 and f_2, we denote by $\pi^*(f_1(\overline{x}), f_2(\overline{x}))$ a coupling between random variables $f_1(\overline{x}), f_2(\overline{x})$ such that

$$\mathscr{M}_{\pi^*}(A) = \begin{cases} \mathscr{M}_{\overline{x}}(B) & \exists B \subset \mathcal{B}(\mathcal{X}) \text{ such that } A = f_1(B) \times f_2(B) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(\mathcal{X})$ is the set of all Borel sets over $\mathcal{X}$, $\mathscr{M}_{\pi^*}(A)$ is the measure that π^* assigns to the Borel set A, and $\mathscr{M}_{\overline{x}}(B)$ is the measure that random variable $\overline{x}$ assigns to Borel set B. Let $\Omega_{\pi^*} = \cup_{\{B|\mathscr{M}_{\overline{x}}(B)\neq 0\}} B$.

C.1 Proof of Theorem 17

Proof. It is easy to verify that $N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\Delta_d}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\mathcal{X}_d})$. Since we know that $\overline{\Delta_d} \subset \overline{\mathcal{X}_d}$, we have

$$N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\Delta_d}) = \sup_{\substack{\overline{S} \subset \overline{\Delta_d} \\ |\overline{S}|=m}} \left\{N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, d_{TV}^{\infty})\right\} \leq \sup_{\substack{\overline{S} \subset \overline{\mathcal{X}_d} \\ |\overline{S}|=m}} \left\{N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, d_{TV}^{\infty})\right\} = N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\mathcal{X}_d}). \tag{5}$$

Let $S = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ be an input set. Denote $\overline{S} = \{\overline{\delta_{x_1}}, \ldots, \overline{\delta_{x_m}}\} \subset \overline{\Delta_d}$ and let $C = \{\overline{\hat{f}_1}_{|\overline{S}}, \ldots, \overline{\hat{f}_r}_{|\overline{S}} \mid \overline{\hat{f}_r} \in \overline{\mathcal{F}}, i \in [r]\}$ be an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S}}$ with respect to d_{TV}^{∞}. Define a new set of non-random functions $\hat{\mathcal{H}} = \left\{\hat{h}_i(x) = \mathbb{E}_{\overline{\hat{f}_i}}\left[\ \overline{\hat{f}_i}(x)\right] \mid i \in [r]\right\}$.

Given any random function $\overline{f} \in \overline{\mathcal{F}}$ and considering the fact that C is an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S}}$ and that $\overline{f}_{|\overline{S}} \in \overline{\mathcal{F}}_{|\overline{S}}$, we know there exists $\overline{\hat{f}_i}$, $i \in [r]$ such that

$$d_{TV}^{\infty}\left(\overline{\hat{f}_i}_{|\overline{S}}, \overline{f}_{|\overline{S}}\right) = d_{TV}^{\infty}\left((\overline{\hat{f}_i}(\overline{\delta_{x_1}}), \ldots, \overline{\hat{f}_i}(\overline{\delta_{x_m}})), (\overline{f}(\overline{\delta_{x_1}}), \ldots, \overline{f}(\overline{\delta_{x_m}}))\right) \leq \epsilon. \tag{6}$$

From Equation 6 we can conclude that for any $k \in [m]$, $d_{TV}\left(\overline{\hat{f}_i}(\overline{\delta_{x_k}}), \overline{f}(\overline{\delta_{x_k}})\right) \leq \epsilon$. Further, for the corresponding $h, \hat{h}_i \in \mathcal{H}$, we know that

$$\hat{h}_i(x_k) = \mathbb{E}_{\overline{\hat{f}_i}}\left[\ \overline{\hat{f}_i}(\overline{\delta_{x_k}})\right] = \int_{\mathbb{R}^d} x \mathscr{D}(\overline{\hat{f}_i}(\overline{\delta_{x_k}}))(x)dx,$$

$$h(x_k) = \mathbb{E}_{\overline{f}}\left[\ \overline{f}(\overline{\delta_{x_k}})\right] = \int_{\mathbb{R}^d} x \mathscr{D}(\overline{f}(\overline{\delta_{x_k}}))(x)dx.$$

Denote $I = \mathscr{D}(\overline{f}(\overline{\delta_{x_k}}))$ and $\hat{I} = \mathscr{D}(\overline{\hat{f}_i}(\overline{\delta_{x_k}}))$. Define two new density functions I_{diff} and $\hat{I}_{diff}$ as

$$I_{diff}(x) = \begin{cases} \dfrac{I(x) - \hat{I}(x)}{d_{TV}(I, \hat{I})} & I(x) \geq \hat{I}(x) \\ 0 & \text{otherwise,} \end{cases}$$

$$\hat{I}_{diff}(x) = \begin{cases} \dfrac{\hat{I}(x) - I(x)}{d_{TV}(I, \hat{I})} & \hat{I}(x) \geq I(x) \\ 0 & \text{otherwise.} \end{cases}$$

Also, we define I_{min} as

$$I_{min}(x) = \frac{\min\{I(x), \hat{I}(x)\}}{\int \min\{I(x), \hat{I}(x)\}dx} = \frac{\min\{I(x), \hat{I}(x)\}}{1 - d_{TV}(I, \hat{I})}.$$

It is easy to verify that

$$I(x) = \left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).I_{diff}(x)$$
$$\hat{I}(x) = \left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).\hat{I}_{diff}(x).$$

We can then find the ℓ_2 distance between $\hat{h}_i(x_k)$ and $h(x_k)$ by

$$\begin{aligned}
&\left\|\hat{h}_i(x_k) - h(x_k)\right\|_2 \\
&= \left\|\int_{\mathbb{R}^d} x\hat{I}(x)dx - \int_{\mathbb{R}^d} xI(x)dx\right\|_2 \\
&= \left\|\int_{\mathbb{R}^d} x\left[\left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).\hat{I}_{diff}(x)\right]\right. \\
&\left. -x\left[\left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).I_{diff}(x)\right] dx\right\|_2 \\
&= \left\|\int_{\mathbb{R}^d} x d_{TV}(I, \hat{I})\left[\hat{I}_{diff}(x) - I_{diff}(x)\right] dx\right\|_2 \\
&= d_{TV}(I, \hat{I})\left\|\int_{\mathbb{R}^d} x\left[\hat{I}_{diff}(x) - I_{diff}(x)\right] dx\right\|_2 \\
&\leq 2B\sqrt{p}\, d_{TV}\left(\overline{f}(\overline{\delta_{x_k}}), \overline{\hat{f}_i}(\overline{\delta_{x_k}})\right) && \text{(Bounded domain } [-B, B]^p \text{ and triangle inequality)} \\
&\leq 2B\epsilon\sqrt{p}.
\end{aligned}$$

Since this result holds for any $k \in [m]$, we have

$$\left\|\hat{h}_{i|S} - h_{|S}\right\|_2^\infty = \left\|(\hat{h}_i(x_1), \ldots, \hat{h}_i(x_m)) - (h(x_1), \ldots, h(x_m))\right\|_2^\infty \leq 2B\epsilon\sqrt{p}. \tag{7}$$

In other words, for any $h_{|S} \in \mathcal{H}_{|S}$ there exists a $\hat{h}_{i|S} \in \hat{\mathcal{H}}_{|S}$ such that $\left\|\hat{h}_{i|S} - h_{|S}\right\|_2^\infty \leq 2B\epsilon\sqrt{p}$.

Therefore, $\hat{\mathcal{H}}_{|S}$ is a $2B\epsilon\sqrt{p}$ cover for $\mathcal{H}_{|S}$ with respect to $\|.\|_2^\infty$ and $|\hat{\mathcal{H}}_{|S}| = r$.

The bound in Equation 7 holds for any subset S of $\mathbb{R}^d$ with $|S| = m$. Therefore,

$$N_U(2B\epsilon\sqrt{p}, \mathcal{H}, m, \|.\|_2^\infty) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\Delta_d}). \tag{8}$$

Putting Equations 5 and 8 together, we conclude

$$N_U(2B\epsilon\sqrt{p}, \mathcal{H}, m, \|.\|_2^\infty) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\Delta_d}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\mathcal{X}_d}).$$

To prove the second part that involves covering number with respect to $\|.\|_2^{\ell_2}$, we can follow the same steps. Similarly, we know that

$$N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) = \sup_{\substack{\overline{S} \subset \overline{\Delta_d} \\ |\overline{S}| = m}} \left\{N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, d_{TV}^{\ell_2})\right\} \leq \sup_{\substack{\overline{S} \subset \overline{\mathcal{X}_d} \\ |\overline{S}| = m}} \left\{N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, d_{TV}^{\ell_2})\right\} = N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\mathcal{X}_d}).$$

Consider the same input sets S and $\overline{S}$ and let and let $\tilde{C} = \{\overline{\tilde{f}}_{1|\overline{S}}, \ldots, \overline{\tilde{f}}_{r|\overline{S}} \mid \overline{\tilde{f}}_t \in \overline{\mathcal{F}}, i \in [t]\}$ be an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S}}$ with respect to $d_{TV}^{\ell_2}$. Define a new set of non-random functions $\tilde{\mathcal{H}} = \left\{\tilde{h}_i(x) = \mathbb{E}_{\overline{\tilde{f}}_i}\left[\overline{\tilde{f}}_i(x)\right] \mid i \in [r]\right\}$.

Similarly, consider $f_{|\overline{S}}$ and $\tilde{f}_{i|\overline{S}}$ such that

$$d_{TV}^{\ell_2}\left(\overline{\tilde{f}}_{i|\overline{S}}, \overline{f}_{|\overline{S}}\right) = d_{TV}^{\ell_2}\left((\overline{\tilde{f}}_i(\overline{\delta_{x_1}}), \ldots, \overline{\tilde{f}}_i(\overline{\delta_{x_m}})), (\overline{f}(\overline{\delta_{x_1}}), \ldots, \overline{f}(\overline{\delta_{x_m}}))\right) \leq \epsilon.$$

Using the same analysis as before, we can conclude that for any $k \in [m]$,

$$\left\| \tilde{h}_i(x_k) - h(x_k) \right\|_2 \leq 2B\sqrt{p}\, d_{TV}\left(\overline{f}(\overline{\delta_{x_k}}), \overline{\tilde{f}_i}(\overline{\delta_{x_k}}) \right).$$

We can then conclude that

$$\begin{aligned}
&\|\tilde{h}_{i|S} - h_{|S}\|_2^{\ell_2} \\
&= \sqrt{\frac{1}{m}\sum_{i=1}^{k} \left\| \tilde{h}_i(x_k) - h(x_k) \right\|_2^2} \\
&\leq \sqrt{\frac{1}{m}\sum_{i=1}^{k} (2B\sqrt{p})^2 \left(d_{TV}\left(\overline{f}(\overline{\delta_{x_k}}), \overline{\tilde{f}_i}(\overline{\delta_{x_k}}) \right) \right)^2} \\
&\leq 2B\sqrt{p} \sqrt{\frac{1}{m}\sum_{i=1}^{k} \left(d_{TV}\left(\overline{f}(\overline{\delta_{x_k}}), \overline{\tilde{f}_i}(\overline{\delta_{x_k}}) \right) \right)^2} \\
&\leq 2B\sqrt{p}\, d_{TV}^{\ell_2}\left(\overline{\tilde{f}_i}_{|\overline{S}}, \overline{f}_{|\overline{S}} \right) \\
&\leq 2B\epsilon\sqrt{p}.
\end{aligned}$$

We can then say that $\tilde{\mathcal{H}}_{|S}$ is a $2B\epsilon\sqrt{p}$ cover for $\mathcal{H}_{|S}$ with respect to $\|.\|_2^{\ell_2}$ and $|\hat{\mathcal{H}}_{|S}| = t$. It follows that

$$N_U(2B\epsilon\sqrt{p}, \mathcal{H}, m, \|.\|_2^{\ell_2}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\mathcal{X}_d}).$$

□

C.2 Proof of Lemma 18

Proof. Denote $\overline{\mathcal{Q}} = \overline{\mathcal{H}} \circ \overline{\mathcal{F}}$. Consider an input set of random variables $\overline{S} = \{\overline{x_1}, \ldots, \overline{x_m}\} \subset \overline{\mathcal{X}_d}$. Denote $r_1 = N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, d_{TV}^{\infty})$ and let $\overline{C} = \{\overline{\hat{f}_1}_{|\overline{S}}, \ldots, \overline{\hat{f}_{r_1}}_{|\overline{S}} \mid \overline{\hat{f}_i} \in \overline{\mathcal{F}}, i \in [r_1]\}$ be an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S}}$ with respect to d_{TV}^{∞} and $\overline{S'} = \{\overline{\hat{f}_i}(\overline{x_k}) \mid i \in [r_1], k \in [m]\}$. Clearly, $|\overline{S'}| \leq mr_1$. Also, let $\overline{C'} = \{\overline{\hat{h}_1}_{|\overline{S'}}, \ldots, \overline{\hat{h}_{r_2}}_{|\overline{S'}} \mid \overline{\hat{h}_j} \in \overline{\mathcal{H}}, j \in [r_2]\}$ be an ϵ'-cover for $\overline{\mathcal{H}}_{|\overline{S'}}$ with respect to d_{TV}^{∞} metric, where $r_2 = N(\epsilon', \overline{\mathcal{H}}_{|\overline{S'}}, d_{TV}^{\infty})$ is the cardinality of the cover set $\overline{C'}$. Denote $\overline{\hat{\mathcal{Q}}} = \{\overline{\hat{h}_j} \circ \overline{\hat{f}_i} \mid i \in [r_1], j \in [r_2]\}$. We claim that $\overline{\hat{\mathcal{Q}}}_{|\overline{S}}$ is an $(\epsilon + \epsilon')$-cover for $\overline{\mathcal{Q}}_{|\overline{S}}$ with respect to d_{TV}^{∞}. Since the cardinality of $\overline{\hat{\mathcal{Q}}}_{|\overline{S}}$ is no more than $r_1 r_2$, we can conclude that $N(\epsilon, \overline{\mathcal{Q}}_{|\overline{S}}, d_{TV}^{\infty}) \leq N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, d_{TV}^{\infty}) N(\epsilon', \overline{\mathcal{H}}_{|\overline{S'}}, d_{TV}^{\infty})$.

Consider $(\overline{h} \circ \overline{f})_{|\overline{S}} = \left(\overline{h}(\overline{f}(\overline{x_1})), \ldots, \overline{h}(\overline{f}(\overline{x_m}))\right) \in \overline{\mathcal{Q}}_{|\overline{S}}$, where $\overline{f} \in \overline{\mathcal{F}}$ and $\overline{h} \in \overline{\mathcal{H}}$. Since $\overline{\mathcal{F}}_{|\overline{S}}$ is ϵ-covered by $\overline{C}$, we know that there exists $\overline{\hat{f}_i} \in \overline{\mathcal{F}}$ such that

$$d_{TV}^{\infty}\left((\overline{\hat{f}_i}(\overline{x_1}), \ldots, \overline{\hat{f}_i}(\overline{x_m})), (\overline{f}(\overline{x_1}), \ldots, \overline{f}(\overline{x_m})) \right) \leq \epsilon.$$

By data processing inequality for total variation distance (Lemma 29), we conclude that $d_{TV}\left(\overline{h}(\overline{\hat{f}_i}(\overline{x_k})), \overline{h}(\overline{f}(\overline{x_k})) \right) \leq \epsilon$ for $k \in [m]$. Therefore,

$$d_{TV}^{\infty}\left((\overline{h}(\overline{\hat{f}_i}(\overline{x_1})), \ldots, \overline{h}(\overline{\hat{f}_i}(\overline{x_m}))), (\overline{h}(\overline{f}(\overline{x_1})), \ldots, \overline{h}(\overline{f}(\overline{x_m}))) \right) \leq \epsilon. \tag{9}$$

Since $\overline{\hat{f}_i}_{|\overline{S}} = (\overline{\hat{f}_i}(\overline{x_1}), \ldots, \overline{\hat{f}_i}(\overline{x_m})) \in \overline{C}$, we know that $\overline{\hat{f}_i}(\overline{x_k}) \in \overline{S'}$ for $k \in [m]$. We also know that $\overline{\mathcal{H}}_{|\overline{S'}}$ is ϵ'-covered by $\overline{C'}$, therefore, there exists $\overline{\hat{h}_j} \in \overline{\mathcal{H}}$ such that

$$d_{TV}^{\infty}\left((\overline{\hat{h}_j}(\overline{\hat{f}_i}(\overline{x_1})), \ldots, \overline{\hat{h}_j}(\overline{\hat{f}_i}(\overline{x_m}))), (\overline{h}(\overline{\hat{f}_i}(\overline{x_1})), \ldots, \overline{h}(\overline{\hat{f}_i}(\overline{x_m}))) \right) \leq \epsilon' \tag{10}$$

Combining Equations 9 and 10 and by using triangle inequality for total variation distance, we conclude that

$$d_{TV}^{\infty}\left((\overline{\hat{h}_j}(\overline{\hat{f}_i}(\overline{x_1})), \ldots, \overline{\hat{h}_j}(\overline{\hat{f}_i}(\overline{x_m}))), \left(\overline{h}(\overline{f}(\overline{x_1})), \ldots, \overline{h}(\overline{f}(\overline{x_m}))\right) \right) \leq \epsilon + \epsilon',$$

which suggests that for any $(\overline{h} \circ \overline{f})_{|\overline{S}} \in \overline{\mathcal{Q}}_{|\overline{S}}$, there exists $(\overline{\hat{h}_j} \circ \overline{\hat{f}_i})_{|\overline{S}} \in \overline{\hat{\mathcal{Q}}}_{|\overline{S}}$ such that

$$d_{TV}^{\infty}\left((\overline{h} \circ \overline{f})_{|\overline{S}}, (\overline{\hat{h}_j} \circ \overline{\hat{f}_i})_{|\overline{S}}\right) \leq \epsilon + \epsilon'.$$

In other words, $\overline{\mathcal{Q}}_{|\overline{S}}$ is $(\epsilon + \epsilon')$-covered by $\overline{\hat{\mathcal{Q}}}_{|\overline{S}}$.

Let $N_1 = N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\mathcal{X}_d})$. We know that $mr_1 \leq mN_1$ and, therefore, $N(\epsilon', \overline{\mathcal{H}}_{|\overline{S'}}, d_{TV}^{\infty}) \leq N_U(\epsilon', \overline{\mathcal{H}}, mr_1, d_{TV}^{\infty}, \overline{\mathcal{X}_d}) \leq N_U(\epsilon', \overline{\mathcal{H}}, mN_1, d_{TV}^{\infty}, \overline{\mathcal{X}_d})$. Since the result holds for any input $\overline{S} \subset \overline{\mathcal{X}_d}$ of cardinality m and we know that $r_1 \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\mathcal{X}_d})$, it follows that

$$N_U(\epsilon + \epsilon', \overline{\mathcal{Q}}, m, d_{TV}^{\infty}, \overline{\mathcal{X}_d}) \leq N_U(\epsilon', \overline{\mathcal{H}}, mN_1, d_{TV}^{\infty}, \overline{\mathcal{X}_d}).N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\mathcal{X}_d}).$$

The bound for $\overline{\Delta_d}$ is almost exactly the same as that of $\overline{\mathcal{X}_d}$. The only difference is that $\overline{S} = \{\overline{\delta_{x_1}}, \ldots, \overline{\delta_{x_m}}\} \subseteq \overline{\Delta_d}$, and we have a uniform ϵ-covering number with respect to $\overline{\Delta_d}$. We conclude that

$$N_U\left(\epsilon + \epsilon', \overline{\mathcal{H}} \circ \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\Delta_d}\right) \leq N_U\left(\epsilon', \overline{\mathcal{H}}, mN_2, d_{TV}^{\infty}, \overline{\mathcal{X}_d}\right).N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\Delta_d}).$$

The bound with respect to $d_{TV}^{\ell_2}$ follows the same analysis. Consider a new set $\overline{S_z} = \{\overline{\delta z_1}, \ldots, \overline{\delta z_m}\} \subset \overline{\Delta_d}$. Denote $t_1 = N(\epsilon, \overline{\mathcal{F}}_{|\overline{S_z}}, d_{TV}^{\ell_2})$ and let $\overline{C_z} = \{\overline{\tilde{f}_1}_{|\overline{S_z}}, \ldots, \overline{\tilde{f}_{t_1}}_{|\overline{S_z}} \mid \overline{\tilde{f}_i} \in \overline{\mathcal{F}}, i \in [t_1]\}$ be an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S_z}}$ with respect to $d_{TV}^{\ell_2}$ and $\overline{S'_z} = \{\overline{\tilde{f}_i}(\overline{\delta z_k}) \mid i \in [t_1], k \in [m]\}$. Clearly, $|\overline{S'_z}| \leq mt_1$. Let $\overline{C'_z} = \{\overline{\tilde{h}_1}_{|\overline{S'_z}}, \ldots, \overline{\tilde{h}_{t_2}}_{|\overline{S'_z}} \mid \overline{\tilde{h}_j} \in \overline{\mathcal{H}}, j \in [t_2]\}$ be an ϵ'-cover for $\overline{\mathcal{H}}_{|\overline{S'_z}}$ with respect to d_{TV}^{∞} metric, where $t_2 = N(\epsilon', \overline{\mathcal{H}}_{|\overline{S'_z}}, d_{TV}^{\infty})$ is the cardinality of the cover set $\overline{C'_z}$. Denote $\overline{\tilde{\mathcal{Q}}} = \{\overline{\tilde{h}_j} \circ \overline{\tilde{f}_i} \mid i \in [t_1], j \in [t_2]\}$. We claim that $\overline{\tilde{\mathcal{Q}}}_{|\overline{S_z}}$ is an $(\epsilon + \epsilon')$-cover for $\overline{\mathcal{Q}}_{|\overline{S_z}}$ with respect to $d_{TV}^{\ell_2}$. We can then conclude that $N(\epsilon, \overline{\mathcal{Q}}_{|\overline{S_z}}, d_{TV}^{\ell_2}) \leq N(\epsilon, \overline{\mathcal{F}}_{|\overline{S_z}}, d_{TV}^{\ell_2}).N(\epsilon', \overline{\mathcal{H}}_{|\overline{S'_z}}, d_{TV}^{\infty})$.

Consider $(\overline{h} \circ \overline{f})_{|\overline{S_z}} = \left(\overline{h}(\overline{f}(\overline{\delta_{z_1}})), \ldots, \overline{h}(\overline{f}(\overline{\delta_{z_m}}))\right) \in \overline{\mathcal{Q}}_{|\overline{S_z}}$, where $\overline{f} \in \overline{\mathcal{F}}$ and $\overline{h} \in \overline{\mathcal{H}}$. Since $\overline{\mathcal{F}}_{|\overline{S_z}}$ is ϵ-covered by $\overline{C_z}$, we know that there exists $\overline{\tilde{f}_i} \in \overline{\mathcal{F}}$ such that

$$\begin{aligned} &d_{TV}^{\ell_2}\left((\overline{\tilde{f}_i}(\overline{\delta_{z_1}}), \ldots, \overline{\tilde{f}_i}(\overline{\delta_{z_m}})), (\overline{f}(\overline{\delta_{z_1}}), \ldots, \overline{f}(\overline{\delta_{z_m}}))\right) \\ &= \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}}), \overline{f}(\overline{\delta_{z_k}}))\right)^2} \leq \epsilon. \end{aligned}$$

Similarly, by data processing inequality, we conclude that $d_{TV}\left(\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}}))\right) \leq d_{TV}\left(\overline{\tilde{f}_i}(\overline{\delta_{z_k}}), \overline{f}(\overline{\delta_{z_k}})\right)$ for $k \in [m]$. Therefore,

$$\begin{aligned} &d_{TV}^{\ell_2}\left((\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_1}})), \ldots, \overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_m}}))), (\overline{hf}(\overline{\delta_{z_1}})), \ldots, \overline{h}(\overline{f}(\overline{\delta_{z_m}})))\right) \\ &= \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}(\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}})))\right)^2} \\ &\leq \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}}), \overline{f}(\overline{\delta_{z_k}}))\right)^2} \leq \epsilon. \end{aligned} \tag{11}$$

Now, using the fact that $\overline{\tilde{f}_i}_{|\overline{S_z}} = (\overline{\tilde{f}_i}(\overline{\delta_{z_1}}), \ldots, \overline{\tilde{f}_i}(\overline{\delta_{z_m}})) \in \overline{C_z}$, we know that $\overline{\tilde{f}_i}(\overline{\delta_{z_k}}) \in \overline{S'_z}$ for $k \in [m]$. We also know that $\overline{\mathcal{H}}_{|\overline{S'_z}}$ is ϵ'-covered by $\overline{C'_z}$ with respect to d_{TV}^{∞}. Therefore, there exists $\overline{\tilde{h}_j} \in \overline{\mathcal{H}}$ such that

$$d_{TV}^{\infty}\left((\overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_1}})), \ldots, \overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_m}}))), (\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_1}})), \ldots, \overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_m}})))\right) \leq \epsilon'. \tag{12}$$

From Equation 12 we can conclude that $d_{TV}\left((\overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), (\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}}))\right) \leq \epsilon'$ for $k \in [m]$. Using triangle inequality for total variation distance, we can write

$$\begin{aligned}
&d_{TV}\left((\overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), (\overline{h}(\overline{f}(\overline{\delta_{z_k}}))\right) \\
&\leq d_{TV}\left((\overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), (\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}}))\right) + d_{TV}\left(\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}}))\right) \\
&\leq d_{TV}\left(\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}}))\right) + \epsilon'.
\end{aligned} \tag{13}$$

We can then conclude that

$$\begin{aligned}
&d_{TV}^{\ell_2}\left((\overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_1}})), \ldots, \overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_m}}))), (\overline{h}(\overline{f}(\overline{\delta_{z_1}})), \ldots, \overline{h}(\overline{f}(\overline{\delta_{z_m}})))\right) \\
&= \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}(\overline{\tilde{h}_j}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}})))\right)^2} \\
&\leq \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}(\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}}))) + \epsilon'\right)^2} && \text{(From Equation 13)} \\
&\leq \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}(\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}})))\right)^2 + \frac{1}{m}\sum_{k=1}^{m}\epsilon'^2} \\
&\leq \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}(\overline{h}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{h}(\overline{f}(\overline{\delta_{z_k}})))\right)^2} + \sqrt{\frac{1}{m}\sum_{k=1}^{m}\epsilon'^2} \\
&\leq \epsilon + \epsilon'. && \text{(From Equation 11)}
\end{aligned}$$

As a result, $\overline{\mathcal{Q}}_{|\overline{S_z}}$ is $(\epsilon+\epsilon')$-covered by $\overline{\tilde{\mathcal{Q}}}_{|\overline{S_z}}$. Let $N_3 = N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d})$. Since $mt_1 \leq mN_3$, we can write $N(\epsilon', \overline{\mathcal{H}}_{|\overline{S'}}, d_{TV}^{\infty}) \leq N_U(\epsilon', \overline{\mathcal{H}}, mt_1, d_{TV}^{\infty}, \overline{\mathcal{X}_d}) \leq N_U(\epsilon', \overline{\mathcal{H}}, mN_3, d_{TV}^{\infty}, \overline{\mathcal{X}_d})$. We know that the result holds for any input $\overline{S_z} \subset \overline{\Delta_d}$ of cardinality m and $t_1 \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d})$, therefore, it follows that

$$N_U(\epsilon+\epsilon', \overline{\mathcal{Q}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) \leq N_U(\epsilon', \overline{\mathcal{H}}, mN_3, d_{TV}^{\infty}, \overline{\mathcal{X}_d}) N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}).$$

□

C.3 TV distance of composition of a class with Gaussian noise class

The following lemma, which bounds the total varaition distance by Wasserstein distance, is borrowed from Chae and Walker (2020).

Lemma 33 (Bounding TV distance by Wasserstein distance)**.** *Given a density function K over $\mathbb{R}^d$ and two probability measures μ, ν over $\mathcal{X}$ with probability density functions I_μ and I_ν, respectively, the total variation distance can be upper bounded in terms of Wasserstein distance between μ and ν.*

$$\|K * I_\mu - K * I_\nu\|_1 \leq \sup_{y \neq z} \frac{\|K(x-y) - K(x-z)\|_1}{\|y-z\|_2} d_{\mathcal{W}}(\mu, \nu)$$

Proof. For any coupling π of μ and ν, we have

$$K * I_\mu(x) - K * I_\nu(x) = \int (K(x-y) - K(x-z)) d\pi(y, z).$$

Therefore,

$$\begin{aligned}
\|K * (I_\mu - I_\nu)\|_1 &= \int \left| \int \left((K(x-y) - K(x-z)) \, d\pi(y,z) \right| dx \\
&\leq \int \int |(K(x-y) - K(x-z)| \, d\pi(y,z) dx && \text{(By Jensen's inequality)} \\
&= \int \|K(x-y) - K(x-z)\|_1 \, d\pi(y,z) && \text{(By Fubini's theorem)} \\
&\leq \sup_{y \neq z} \left\{ \frac{\|K(x-y) - K(x-z)\|_1}{\|y-z\|_2} \right\} \int \|y-z\|_2 d\pi(y,z)
\end{aligned}$$

Since this holds for any coupling π of μ and ν we conclude that

$$\|K * (I_\mu - I_\nu)\|_1 \leq \sup_{y \neq z} \left\{ \frac{\|K(x-y) - K(x-z)\|_1}{\|y-z\|_2} \right\} d_{\mathcal{W}}(\mu, \nu)$$

□

C.4 Proof of Theorem 20

Proof. Fix an input set $\overline{S} = \{\overline{x_1}, \ldots, \overline{x_m}\} \subset \mathbb{R}^d$. Let $\overline{C} = \{\overline{\hat{f}_1}_{|\overline{S}}, \ldots, \overline{\hat{f}_r}_{|\overline{S}} : \overline{\hat{f}_i} \in \overline{\mathcal{F}}, i \in [r]\}$ be an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S}}$ with respect to $d_{\mathcal{W}}^\infty$ metric. Denote $\overline{\mathcal{Q}} = \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}$.We define a new class of random functions $\overline{\hat{\mathcal{Q}}} = \{\overline{g_\sigma} \circ \overline{\hat{f}_i} \mid i \in [r]\}$. We show that $\overline{\mathcal{Q}}_{|\overline{S}}$ is $(\frac{9\epsilon}{2\sigma})$-covered by $\overline{\hat{\mathcal{Q}}}_{|\overline{S}}$ and since $|\overline{\hat{Q}}_{|\overline{S}}| = r$, the result follows.

Let I_σ denote the probability density function of $\mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. For any $\overline{f} \in \overline{\mathcal{F}}$, we have $\overline{g_\sigma}(\overline{f}(x)) = \overline{f}(x) + \overline{z}$, where $\overline{z}$ is a random variable with probability density function I_σ, therefore, we know that $\mathscr{D}(\overline{g_\sigma}(\overline{f}(x)) = \mathscr{D}(\overline{f}(x)) * I_\sigma$.

Given $(\overline{g_\sigma} \circ \overline{f})_{|\overline{S}} = (\overline{g_\sigma}(\overline{f}(\overline{x_1})), \ldots, \overline{g_\sigma}(\overline{f}(\overline{x_m}))) \in \overline{\mathcal{Q}}_{|\overline{S}}$, we know that $\overline{f}_{|\overline{S}} = (\overline{f}(\overline{x_1}), \ldots, \overline{f}(\overline{x_m}))$ is in set $\overline{\mathcal{F}}_{|\overline{S}}$. Therefore, there exists $\overline{\hat{f}_i} \in \overline{\mathcal{F}}$ such that $d_{\mathcal{W}}^\infty(\overline{\hat{f}_i}_{|\overline{S}}, \overline{f}_{|\overline{S}}) \leq \epsilon$, i.e.,

$$d_{\mathcal{W}}^\infty \left((\overline{\hat{f}_i}(\overline{x_1}), \ldots, \overline{\hat{f}_i}(\overline{x_m})), (\overline{f}(\overline{x_1}), \ldots, \overline{f}\overline{x_m})) \right) \leq \epsilon. \tag{14}$$

From Equation 14, we know that $d_{\mathcal{W}}(\overline{\hat{f}_i}(\overline{x_k}), \overline{f}(\overline{x_k})) \leq \epsilon$ for all $k \in [m]$. From Lemma 33, we can conclude that

$$\begin{aligned}
&\frac{1}{2} \left\| I_\sigma * \mathscr{D}(\overline{\hat{f}_i}(\overline{x_k}) - I_\sigma * \mathscr{D}(\overline{f}(\overline{x_k}))) \right\|_1 \\
&\leq \frac{1}{2} \left(\sup_{y \neq z} \left\{ \frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2} \right\} \right) d_{\mathcal{W}}(\overline{\hat{f}_i}(\overline{x_k}), \overline{f}\overline{x_k})) \\
&\leq \frac{\epsilon}{2} \left(\sup_{y \neq z} \left\{ \frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2} \right\} \right).
\end{aligned} \tag{15}$$

Moreover, $I_\sigma * \mathscr{D}(\overline{\hat{f}_i}(\overline{x_k})$ and $I_\sigma * \mathscr{D}(\overline{f}(\overline{x_k}))$ are probability density functions of $\overline{g_\sigma}(\overline{\hat{f}_i}(\overline{x_k}))$ and $\overline{g_\sigma}(\overline{f}(\overline{x_k}))$, respectively. Therefore, from Equation 15,

$$d_{TV} \left(\overline{g_\sigma}(\overline{\hat{f}_i}(\overline{x_k})), \overline{g_\sigma}(\overline{f}(\overline{x_k})) \right) \leq \frac{\epsilon}{2} \left(\sup_{y \neq z} \left\{ \frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2} \right\} \right). \tag{16}$$

Since Equation 16 holds for all $k \in [m]$, it follows that

$$d_{TV}^\infty \left(\overline{g_\sigma}(\overline{\hat{f}_i}(\overline{x_k})), \overline{g_\sigma}(\overline{f}(\overline{x_k})) \right) \leq \frac{\epsilon}{2} \left(\sup_{y \neq z} \left\{ \frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2} \right\} \right).$$

This shows that for any $(\overline{g_\sigma} \circ \overline{f})_{|\overline{S}} \in \overline{\mathcal{Q}}_{|\overline{S}}$ there exists $(\overline{g_\sigma} \circ \overline{\hat{f}_i})_{|\overline{S}} \in \overline{\hat{\mathcal{Q}}}_{|\overline{S}}$ such that

$$d_{TV}^{\infty}\left((\overline{g_\sigma} \circ \overline{f})_{|\overline{S}}, (\overline{g_\sigma} \circ \overline{\hat{f}_i})_{|\overline{S}}\right) \leq \frac{\epsilon}{2}\left(\sup_{y \neq z}\left\{\frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2}\right\}\right). \tag{17}$$

It is only left to bound the supremum term in Equation 17.

Based on Theorem 30, we know that for two Gaussian distributions $\mathcal{N}(\mu_1, \sigma^2 I)$ and $\mathcal{N}(\mu_2, \sigma^2 I)$ their total variation distance can be bounded by

$$d_{TV}\left(\mathcal{N}(\mu_1, \sigma^2 I), \mathcal{N}(\mu_2, \sigma^2 I)\right) \leq \frac{9}{2}\min\left\{1, \frac{\|\mu_1 - \mu_2\|_2}{\sigma}\right\}. \tag{18}$$

We also know that $\|I_\sigma(x-y) - I_\sigma(x-z)\|_1 = 2d_{TV}(\mathcal{N}(y, \sigma^2 I), \mathcal{N}(z, \sigma^2 I))$. Combining Equations 17 and 18, we can write

$$d_{TV}^{\infty}\left((\overline{g_\sigma} \circ \overline{f})_{|\overline{S}}, (\overline{g_\sigma} \circ \overline{\hat{f}_i})_{|\overline{S}}\right) \leq \frac{\epsilon}{2}\left(\sup_{y \neq z}\left\{9\frac{\min\left\{1, \frac{\|y-z\|_2}{\sigma}\right\}}{\|y-z\|_2}\right\}\right) \leq \frac{9}{2}\frac{\epsilon}{\sigma}. \tag{19}$$

From Equation 19 it follows that $\overline{\mathcal{Q}}_{|\overline{S}}$ is $(\frac{9\epsilon}{2\sigma})$-covered by $\overline{\hat{\mathcal{Q}}}_{|\overline{S}}$. Since the result holds for any subset $\overline{S}$ of $\overline{\mathcal{X}_d}$ with cardinality m, we can conclude that

$$N_U\left(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\mathcal{X}_d}\right) \leq N_U(\epsilon, \mathcal{F}, m, d_{\mathcal{W}}^{\infty}, \overline{\mathcal{X}_d}).$$

The second part of the proof is similar. We consider a set of inputs $\overline{S_z} = \{\overline{\delta_{z_1}}, \ldots, \overline{\delta_{z_m}}\} \subset \overline{\Delta_d}$. We can then consider an ϵ-cover $\overline{C_z} = \{\overline{\tilde{f}}_{1|\overline{S}}, \ldots, \overline{\tilde{f}}_{t|\overline{S}} : \overline{\tilde{f}_i} \in \overline{\mathcal{F}}, i \in [t]\}$ for $\overline{\mathcal{F}}_{|\overline{S_z}}$. We will then construct a class of functions $\overline{\tilde{\mathcal{Q}}} = \{\overline{g_\sigma} \circ \overline{\tilde{f}_i} \mid i \in [t]\}$ and show that $\overline{\mathcal{Q}}_{|\overline{S_z}}$ is $(\frac{9\epsilon}{2\sigma})$-covered by $\overline{\tilde{\mathcal{Q}}}_{|\overline{S_z}}$. The proof follows the same steps as the previous part. Particularly, let $\overline{\tilde{f}_i} \in \overline{\mathcal{F}}$ be such that $d_{\mathcal{W}}^{\infty}(\overline{\tilde{f}_i}_{|\overline{S_z}}, \overline{f}_{|\overline{S_z}}) \leq \epsilon$. For any $k \in [m]$, we can write that

$$\begin{aligned}
&\frac{1}{2}\left\|I_\sigma * \mathscr{D}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}}) - I_\sigma * \mathscr{D}(\overline{f}(\overline{\delta_{z_k}}))\right\|_1 \\
&\leq \frac{1}{2}\left(\sup_{y \neq z}\left\{\frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2}\right\}\right) d_{\mathcal{W}}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}}), \overline{f}(\overline{\delta_{z_k}})) \\
&\leq \frac{\epsilon}{2}\left(\sup_{y \neq z}\left\{\frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2}\right\}\right).
\end{aligned} \tag{20}$$

Using the same arguments as the previous part, we will have that

$$d_{TV}\left(\overline{g_\sigma}(\overline{\tilde{f}_i}(\overline{\delta_{z_k}})), \overline{g_\sigma}(\overline{f}(\overline{\delta_{z_k}}))\right) \leq \frac{\epsilon}{2}\left(\sup_{y \neq z}\left\{\frac{\|I_\sigma(x-y) - I_\sigma(x-z)\|_1}{\|y-z\|_2}\right\}\right) \leq \frac{9}{2}\frac{\epsilon}{\sigma}.$$

Therefore, we can conclude that for any $\overline{f} \in \overline{\mathcal{F}}$ there exists $\overline{\tilde{f}_i}$, $i \in [t]$ such that

$$d_{TV}^{\infty}\left((\overline{g_\sigma} \circ \overline{f})_{|\overline{S_z}}, (\overline{g_\sigma} \circ \overline{\tilde{f}_i})_{|\overline{S_z}}\right) \leq \frac{9}{2}\frac{\epsilon}{\sigma},$$

which means that $\overline{\mathcal{Q}}_{\overline{S_z}}$ is $(\frac{9\epsilon}{2\sigma})$-covered by $\overline{\tilde{\mathcal{Q}}}_{|\overline{S_z}}$. Since the result holds for every $\overline{S_z} \subset \overline{\Delta_d}$ of cardinality m, we can conclude that

$$N_U\left(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\Delta_d}\right) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{\mathcal{W}}^{\infty}, \overline{\Delta_d}).$$

□

C.5 Proof of Corollary 21

Proof. First, from Proposition 16, we can conclude that

$$N_U(\epsilon, \mathcal{F}, d_{\mathcal{W}}^{\infty}, m, \overline{\Delta_d}) = N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\infty}). \tag{21}$$

Then, consider an input set $\overline{S_z} = \{\overline{\delta_{x_1}}, \ldots, \overline{\delta_{x_m}}\} \subset \overline{\Delta_d}$. Let $\overline{C_z} = \{\hat{f}_{1|\overline{S_z}}, \ldots, \hat{f}_{r|\overline{S_z}} \mid \hat{f}_i \in \mathcal{F}, i \in [r]\}$ be an ϵ-cover for $\mathcal{F}_{|\overline{S_z}}$ with respect to $d_{\mathcal{W}}^{\infty}$, then for a given $f_{|\overline{S_z}} \in \mathcal{F}_{|\overline{S_z}}$ and $\hat{f}_{i|\overline{S_z}} \in \overline{C_z}$, where $d_{\mathcal{W}}^{\infty}(f_{|\overline{S_z}}, \hat{f}_{i|\overline{S_z}}) \leq \epsilon$, from Equations 15 and 19, we know that for all $k \in [m]$

$$\begin{aligned}
&d_{TV}(\overline{g_\sigma}(\hat{f}_i(\overline{\delta_{x_k}})), \overline{g_\sigma}(f(\overline{\delta_{x_k}}))) = d_{TV}\left(\mathcal{N}(\hat{f}_i(x_k), \sigma^2 I_p), \mathcal{N}(f(x_k), \sigma^2 I_p)\right) \\
&\leq \frac{9}{2}\min\left\{1, \frac{\|\hat{f}_i(x_k) - f(x_k)\|_2}{\sigma}\right\} \leq \frac{9}{2}\frac{d_{\mathcal{W}}\left(\hat{f}_i(\overline{\delta_{x_k}}), f(\overline{\delta_{x_k}})\right)}{\sigma} \\
&\leq \frac{9\epsilon}{2\sigma}.
\end{aligned}$$

Therefore, we can conclude that

$$\begin{aligned}
&d_{TV}^{\infty}\left((\overline{g_\sigma} \circ \hat{f}_i)_{|\overline{S_z}}, (\overline{g_\sigma} \circ f)_{|\overline{S_z}}\right) \\
&= d_{TV}^{\infty}\left((\overline{g_\sigma}(\hat{f}_i(\overline{\delta_{x_1}})), \ldots, \overline{g_\sigma}(\hat{f}_i(\overline{\delta_{x_m}}))), (\overline{g_\sigma}(f(\overline{\delta_{x_1}})), \ldots, \overline{g_\sigma}(f(\overline{\delta_{x_m}})))\right) \\
&\leq \frac{9}{2}\frac{\epsilon}{\sigma},
\end{aligned}$$

It follows that for any $(\overline{g_\sigma} \circ f)_{|\overline{S_z}} \in (\overline{\mathcal{G}_\sigma} \circ \mathcal{F})_{|\overline{S_z}}$, there exists $\hat{f}_{i|\overline{S_z}} \in \overline{C_z}$ such that $d_{TV}^{\infty}\left((\overline{g_\sigma} \circ \hat{f}_i)_{|\overline{S_z}}, (\overline{g_\sigma} \circ f)_{|\overline{S_z}}\right) \leq \frac{9\epsilon}{2\sigma}$. Therefore,

$$N(\frac{9\epsilon}{2\sigma}, (\overline{\mathcal{G}_\sigma} \circ \mathcal{F})_{|\overline{S_z}}, d_{TV}^{\infty}) \leq N(\epsilon, \mathcal{F}_{|\overline{S_z}}, d_{\mathcal{W}}^{\infty}).$$

Since this results holds for any $\overline{S_z} \subset \overline{\Delta_d}$, we can conclude that

$$N_U(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_\sigma} \circ \mathcal{F}, m, d_{TV}^{\infty}, \overline{\Delta_d}) \leq N_U(\epsilon, \mathcal{F}, m, d_{\mathcal{W}}^{\infty}, \overline{\Delta_d}) = N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\infty}).$$

The proof of the second part again follows from Proposition 16. We can write that

$$N_U(\epsilon, \mathcal{F}, d_{\mathcal{W}}^{\ell_2}, m, \overline{\Delta_d}) = N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2}).$$

Consider the input set $\overline{S_z} \subset \overline{\Delta_d}$ as defined above and let $\overline{\tilde{C}_z} = \{\tilde{f}_{1|\overline{S_z}}, \ldots, \tilde{f}_{t|\overline{S_z}} \mid \tilde{f}_i \in \mathcal{F}, i \in [t]\}$ be an ϵ-cover for $\mathcal{F}_{|\overline{S_z}}$ with respect to $d_{\mathcal{W}}^{\ell_2}$. Now, for a given $f_{|\overline{S_z}} \in \mathcal{F}_{|\overline{S_z}}$ and the corresponding $\tilde{f}_{i|\overline{S_z}} \in \overline{\tilde{C}_z}$, where $d_{\mathcal{W}}^{\ell_2}(f_{|\overline{S_z}}, \tilde{f}_{i|\overline{S_z}}) \leq \epsilon$, we know that for all $k \in [m]$

$$\begin{aligned}
&d_{TV}(\overline{g_\sigma}(\tilde{f}_i(\overline{\delta_{x_k}})), \overline{g_\sigma}(f(\overline{\delta_{x_k}}))) = d_{TV}\left(\mathcal{N}(\tilde{f}_i(x_k), \sigma^2 I_p), \mathcal{N}(f(x_k), \sigma^2 I_p)\right) \\
&\leq \frac{9}{2}\min\left\{1, \frac{\|\tilde{f}_i(x_k) - f(x_k)\|_2}{\sigma}\right\} \leq \frac{9}{2}\frac{d_{\mathcal{W}}\left(\tilde{f}_i(\overline{\delta_{x_k}}), f(\overline{\delta_{x_k}})\right)}{\sigma}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&d_{TV}^{\ell_2}\left((\overline{g_\sigma} \circ \tilde{f}_i)_{|\overline{S_z}}, (\overline{g_\sigma} \circ f)_{|\overline{S_z}}\right) \\
&= \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(d_{TV}\left(\overline{g_\sigma}(\tilde{f}_i(\overline{\delta_{x_k}})), \overline{g_\sigma}(f(\overline{\delta_{x_k}}))\right)\right)^2} \\
&\leq \sqrt{\frac{1}{m}\sum_{k=1}^{m}\frac{\left(9\, d_{\mathcal{W}}\left(\tilde{f}_i(\overline{\delta_{x_k}}), f(\overline{\delta_{x_k}})\right)\right)^2}{(2\sigma)^2}} \\
&\leq \frac{9}{2\sigma} d_{\mathcal{W}}^{\ell_2}(\tilde{f}_{i|S_z}, f_{|S_z}) \leq \frac{9}{2}\frac{\epsilon}{\sigma}.
\end{aligned}$$

Therefore, for any $(\overline{g_\sigma} \circ f)_{|\overline{S_z}} \in (\overline{\mathcal{G}_\sigma} \circ \mathcal{F})_{|\overline{S_z}}$, there exists $\tilde{f}_{i|\overline{S_z}} \in \overline{C_z}$ such that $d_{TV}^{\ell_2}\left((\overline{g_\sigma} \circ \tilde{f}_i)_{|\overline{S_z}}, (\overline{g_\sigma} \circ f)_{|\overline{S_z}}\right) \leq \frac{9\epsilon}{2\sigma}$. As a result,

$$N(\frac{9\epsilon}{2\sigma}, (\overline{\mathcal{G}_\sigma} \circ \mathcal{F})_{|S_z}, d_{TV}^{\ell_2}) \leq N(\epsilon, \mathcal{F}_{|S_z}, d_{\mathcal{W}}^{\ell_2}).$$

Since this results holds for any $\overline{S_z} \subset \overline{\Delta_d}$, we can conclude that

$$N_U(\frac{9\epsilon}{2\sigma}, \overline{\mathcal{G}_\sigma} \circ \mathcal{F}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) \leq N_U(\epsilon, \mathcal{F}, m, d_{\mathcal{W}}^{\ell_2}, \overline{\Delta_d}) = N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2}).$$

□

C.6 Proof of Theorem 22

Proof. Let $\overline{\mathcal{Q}} = \overline{\mathcal{G}_\sigma} \circ \mathcal{F}$ and $\overline{\mathcal{Z}} = \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}$. Denote by $r = N_U(\epsilon, \mathcal{F}, \infty, \|.\|_2^\infty)$. Let $C = \{\hat{f}_i(x) \mid \hat{f}_i \in \mathcal{F}, \forall x \in \mathbb{R}^d, i \in [r]\}$ be a global ϵ-cover for $\mathcal{F}$ with respect to $\|.\|_2$ metric. We will show that for all $(\overline{g_\sigma} \circ f)_{|\overline{\mathcal{Z}}}$, $f \in \mathcal{F}$, there exists $\hat{f}_i \in C$ such that $d_{TV}^\infty\left((\overline{g_\sigma} \circ f)_{|\overline{\mathcal{Z}}}, (\overline{g_\sigma} \circ \hat{f}_i)_{|\overline{\mathcal{Z}}}\right) \leq \frac{9\epsilon}{2\sigma}$. Clearly, $|C| \leq r$ and the result follows.

For any $\overline{x} \in \overline{\mathcal{X}_{B,d}}$ and its smoothed version $\overline{z} \in \overline{\mathcal{Z}}$ with probability density function I_z, from Lemma 56, we know that we can estimate I_z by a mixture $\overline{h} = \sum_{i=1}^m w_i g_i$ of $m = \lceil \frac{B}{\eta} \rceil^d$ Gaussian random variables with covariance matrix $\sigma^2 I_d$ and means as defined in Equation 43 such that $d_{TV}(\overline{h}, \overline{z}) \leq 18\sqrt{d}\eta/\sigma$. Let $\overline{\mathcal{H}}$ denote the set containing all mixtures of this kind.

Since C covers the restriction of $\mathcal{F}$ to $\mathbb{R}^d$, for any $f \in \mathcal{F}$, there exists $\hat{f}_i$ such that $\|f(x) - \hat{f}_i(x))\|_2 \leq \epsilon$ for every $x \in \mathbb{R}^d$. Next, for the coupling $\pi^*(f(\overline{h}), \hat{f}_i(\overline{h}))$ as defined in Notations we can write

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2 d\pi^*(x, y) \leq \epsilon \int_{\mathbb{R}^d \times \mathbb{R}^d} d\pi^*(x, y) \leq \epsilon,$$

which comes from the fact that $\hat{f}_i$ is "globally close" to f with respect to $\|.\|_2$ distance. We, therefore, know that

$$d_{\mathcal{W}}(f(\overline{h}), \hat{f}_i(\overline{h})) = \inf_{\pi \in \Pi(f(\overline{h}), \hat{f}_i(\overline{h}))} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2 d\pi(x, y) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2 d\pi^*(x, y) \leq \epsilon.$$

Since this holds for any $\overline{h} \in \overline{\mathcal{H}}$, we can conclude that

$$d_{\mathcal{W}}^\infty\left(f_{|\overline{\mathcal{H}}}, \hat{f}_{i|\overline{\mathcal{H}}}\right) \leq \epsilon.$$

Next, from the arguments in Theorem 20, we know that

$$d_{TV}^\infty\left((\overline{g}_\sigma \circ f)_{|\overline{\mathcal{H}}}, (\overline{g}_\sigma \circ \hat{f}_i)_{|\overline{\mathcal{H}}}\right) \leq \frac{9}{2\sigma} d_{\mathcal{W}}^\infty\left(f_{|\overline{\mathcal{H}}}, \hat{f}_{i|\overline{\mathcal{H}}}\right) \leq \frac{9\epsilon}{2\sigma}.$$

We can now say that for any $\overline{z} \in \overline{\mathcal{Z}}$ and the $\overline{h} \in \overline{\mathcal{H}}$ that estimates it with respect to total variation distance, we have

$$\begin{aligned}
&d_{TV}\left(\overline{g}_\sigma(f(\overline{z})), \overline{g}_\sigma(\hat{f}_i(\overline{z}))\right) \\
&\leq d_{TV}\left(\overline{g}_\sigma(f(\overline{z})), \overline{g}_\sigma(f(\overline{h}))\right) + d_{TV}\left(\overline{g}_\sigma(f(\overline{h})), \overline{g}_\sigma(\hat{f}_i(\overline{h}))\right) + d_{TV}\left(\overline{g}_\sigma(\hat{f}_i(\overline{h})), \overline{g}_\sigma(\hat{f}_i(\overline{z}))\right) \\
&\leq 36\sqrt{d}\eta/\sigma + \frac{9\epsilon}{2\sigma}.
\end{aligned}$$

where we used triangle and data processing inequalities for total variation distance. Setting $\eta = \epsilon/8\sqrt{d}$ we have $d_{TV}\left(\overline{g}_\sigma(f(\overline{z})), \overline{g}_\sigma(\hat{f}_i(\overline{z}))\right) \leq 9\epsilon/\sigma$.

Since we have this result for every $\overline{h} \in \overline{\mathcal{H}}$, we know that for any $\overline{z} \in \overline{\mathcal{Z}}$,

$$d_{TV}^\infty\left((\overline{g}_\sigma \circ f)_{|\overline{\mathcal{Z}}}, (\overline{g}_\sigma \circ \hat{f}_{i|\overline{\mathcal{Z}}})\right) \leq \frac{9\epsilon}{\sigma},$$

which is exactly what we wanted to prove. Therefore, the size of the TV cover for $\overline{\mathcal{G}_\sigma} \circ \mathcal{F}$ can be bounded by the size of $\|.\|_2$ cover of $\mathcal{F}$

$$N_U(\frac{9\epsilon}{\sigma}, \overline{\mathcal{G}_\sigma} \circ \mathcal{F}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}) \leq N_U(\epsilon, \mathcal{F}, \infty, \|.\|_2^{\infty}).$$

□

D Proofs of lemmas in Section 6

Notation. For a vector $V \in \mathbb{R}^d$, we denote its angle by $\angle V$. By $\angle(V_1, V_2)$, we are referring to the angle between two vectors V_1 and V_2. Also, we denote by $1\{x = a\}$ the indicator function that outputs 1 if $x = a$ and 0 if $x \neq a$. We also denote by $\langle V_1, V_2 \rangle$ the inner product between between vectors V_1 and V_2. We denote by $\mathscr{D}(\overline{x})$ the probability density functions of the random variable $\overline{x}$. For two Borel functions f_1 and f_2, we denote by $\pi^*(f_1(\overline{x}), f_2(\overline{x}))$ a coupling between random variables $f_1(\overline{x}), f_2(\overline{x})$ such that

$$\mathscr{M}_{\pi^*}(A) = \begin{cases} \mathscr{M}_{\overline{x}}(B) & \exists B \subset \mathcal{B}(\mathcal{X}) \text{ such that } A = f_1(B) \times f_2(B) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(\mathcal{X})$ is the set of all Borel sets over $\mathcal{X}$, $\mathscr{M}_{\pi^*}(A)$ is the measure that π^* assigns to the Borel set A, and $\mathscr{M}_{\overline{x}}(B)$ is the measure that random variable $\overline{x}$ assigns to Borel set B. We also denote by $Ball_d(x, R)$ the d dimensional ball of radius R centered at x.

D.1 Proof of Theorem 25

In the following we state a stronger version of Theorem 25 which presents a uniform covering number bound for neural network classes that have a general activation function that is Lipschitz continuous, monotone, and has a bounded domain.

Theorem 34 (Stronger version of Theorem 25)**.** *Consider the class NET[d, p] of a single-layers neural network, where the activation function is Lipschitz continuous with Lipschtiz factor L, monotone, and has a bounded output in $[-B, B]^p$. The global covering number of $\overline{\mathcal{G}_\sigma} \circ$ NET[d, p] with respect to total variation distance is bounded by*

$$\begin{aligned} &N_U(\epsilon, \overline{\mathcal{G}_\sigma} \circ \textit{NET[}d, p\textit{]}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}) \\ &\leq \left(\frac{10\sqrt{10}(4+B)^{3/2}}{(2\pi)^{1/4}} \frac{d^{5/2} L \sqrt{Bu}}{\epsilon^{3/2}\sigma^2} \ln\left(\frac{5(4+B)Bd}{\epsilon\sigma}\right) \right)^{p(d+1)}, \end{aligned}$$

where $u = \max\left\{ \left|\phi^{-1}\left(B - \sigma\epsilon/(5(4+B)d)\right)\right|, \left|\phi^{-1}\left(-B + \sigma\epsilon/(5(4+B)d)\right)\right| \right\}$.

Note that Theorem 25 is a special case of the above theorem where the activation function is the sigmoid function with Lipschitz continuity factor of 1 and a bounded domain in $[0, 1]^p$. In the case of sigmoid function, we can also conclude that

$$\begin{aligned} u &= \max\left\{ \left|\phi^{-1}\left(1 - \epsilon\sigma/(5(4+B)d)\right)\right|, \left|\phi^{-1}\left(\epsilon\sigma/(5(4+B)d)\right)\right| \right\} \\ &= \left|\phi^{-1}\left(1 - \epsilon\sigma/(5(4+B)d)\right)\right| \\ &= \ln\left((5(4+B)d - \epsilon\sigma)/(\epsilon\sigma)\right) \\ &\leq \ln\left((30d - \epsilon\sigma)/(\epsilon\sigma)\right). \end{aligned}$$

Proof. We bound the global covering number of class NET[d, p] $= \{f : \mathbb{R}^d \to \mathbb{R}^p \mid f(x) = \Phi(W^\top x)\}$ with respect to Wasserstein distance by constructing a grid for the weights $V_i \in \mathbb{R}^d$ in $W^\top = [V_1^\top \dots V_p^\top]$. Then, we find the TV covering number using Theorem 20. To do so, we consider two cases for each V_i based on its ℓ_2 norm. In case $\|V_i\|_2 \leq B_v$, we construct the grid based on $\|V_i\|_2$ and its angle, while for the case that $\|V_i\|_2 > B_v$, we prove that only a grid on the angle of V_i is sufficient. Further, we choose B_v based on ϵ and σ. We then show that for each matrix $W^\top = [V_1^\top \dots V_p^\top]$, there exists $\hat{W}^\top = [\hat{V}_1^\top \dots \hat{V}_p^\top]$ in the grid such that $d_{\mathcal{W}}\left(\Phi(W^\top \overline{x}), \Phi(\hat{W}^\top \overline{x})\right)$ is bounded for all $\overline{x} \in \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}$.

Denote $r = \lceil \frac{2B_v}{\delta} \rceil$ and

$$A = \{-B_v + i\delta \mid\in [r]\}^d. \tag{22}$$

Define a new set

$$A_S = \left\{ (a_1, \ldots, a_d) \in A \mid \left(\sum_{i=1}^{d} 1\{a_i = B_v\} + \sum_{i=1}^{d} 1\{a_i = -B_v\} \right) \geq 1 \right\}.$$

Informally, A_S is the grid of points on sides of a d-dimensional hypercube. For any point $b = (b_1, \ldots, b_d) \in A_S$, we define the following set of vectors

$$P_b = \{ \frac{i\zeta}{B_v} [b_1 \ldots b_d] \in \mathbb{R}^d \mid i \in [\lceil \frac{B_v}{\zeta} \rceil] \}.$$

Note that the way we defined A_S in Equation 22, implies that for any $(b_1, \ldots, b_d) \in A_S$, there exists at least one b_i such that $|b_i| = B_v$. Therefore, whenever $i = \lceil \frac{B_v}{\zeta} \rceil$, we know that $\| \frac{i\zeta}{B_v} [b_1 \ldots b_d] \|_2 \geq B_v$.

Now, we can define the grid of vectors $V \in \mathbb{R}^d$ in the following way

$$C = \bigcup_{b \in A_S} P_b.$$

Informally speaking, we are discretizing the norms in $\lceil \frac{B_v}{\zeta} \rceil$ values and then for each vector from origin to gird points on the sides of the hypercube, we use $\lceil \frac{B_v}{\zeta} \rceil$ vectors with the same angle and different norms as our grid. Clearly, the size of grid $|C|$ is upper bounded by $\lceil \frac{B_v}{\zeta} \rceil \lceil \frac{2B_v}{\delta} \rceil^d$.

Next, we turn into proving that given any vector V in $\mathbb{R}^d$, there exists a vector $\hat{V}$ in C such that for any $\overline{z} \in \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}$, $d_{\mathcal{W}}(\Phi(V^\top \overline{z}), \Phi(\hat{V}^\top \overline{z})) \leq (3 + 2\sqrt{2})\epsilon$.

Case 1. In this case, we consider vectors $V \in \mathbb{R}^d$ such that $\|V\|_2 \leq B_v$. The way that we constructed the set of vectors C implies that given any vector there exists a $b \in A_S$ and the set of aligned vectors P_b such that the angle between V and vectors in set P_b can be bounded. More specifically, for any $V' \in P_b$, we know that

$$\angle(V, V') \leq \arcsin \frac{\delta}{B_v}.$$

since $\arcsin$ is a monotone increasing functions over $[-1, 1]$ and we know that $\|[b_1 \ldots b_d]\|_2 \geq B_v$. Let $\theta = \arcsin \frac{\delta}{B_v}$. Moreover, since $\|V\|_2 \leq B_v$, we know that there exists $\hat{V} \in P_b$ such that

$$\left| \|V\|_2 - \|\hat{V}\|_2 \right| \leq \frac{\zeta}{B_v} \|[b_1 \ldots b_d]\|_2 \leq \frac{\zeta}{B_v} \sqrt{d} B_v \leq \sqrt{d}\zeta.$$

Without loss of generality, let $\|V\|_2 \leq \|\hat{V}\|_2$. We can then write

$$\frac{\|\hat{V}\|_2}{\|V\|_2} \leq 1 + \frac{\sqrt{d}\zeta}{\|V\|_2}$$

Denote $\hat{V}_\perp = \|\hat{V}\|_2 sin(\angle(V, \hat{V})) V_\perp$ and $\hat{V}_\parallel = \|\hat{V}\|_2 cos(\angle(V, \hat{V})) \frac{V}{\|V\|_2}$, where $V_\perp$ is a normalized vector orthogonal to V. Denote $B_z = (B + \sigma)\sqrt{d} + \sigma\sqrt{2 \ln \frac{B}{\epsilon}}$. For any $x \in \mathbb{R}^d$ such that $\|x\|_2 \leq B_z$,

we can write

$$\begin{aligned}
\langle \hat{V}, x\rangle &= \langle \hat{V}_\perp, x\rangle + \langle \hat{V}_\parallel, x\rangle = \langle \hat{V}_\perp, x\rangle + \langle V, x\rangle \frac{\|\hat{V}_\parallel\|_2}{\|V\|_2} \\
&= \|\hat{V}_\perp\|_2 \|x\|_2 cos(\angle(\hat{V}_\perp, x)) + \langle V, x\rangle \frac{\|\hat{V}_\parallel\|_2}{\|V\|_2} \\
&\leq \|\hat{V}_\perp\|_2 \|x\|_2 + \langle V, x\rangle \frac{\|\hat{V}_\parallel\|_2}{\|V\|_2} \\
&\leq \|\hat{V}\|_2 \|x\|_2 sin(\angle(V, \hat{V})) + \langle V, x\rangle \frac{\|\hat{V}\|_2 cos(\angle(V, \hat{V}))}{\|V\|_2} \\
&\leq \|\hat{V}\|_2 \|x\|_2 \frac{\delta}{B_v} + \langle V, x\rangle \frac{\|\hat{V}\|_2}{\|V\|_2} \\
&\leq \sqrt{d} B_v \|x\|_2 \frac{\delta}{B_v} + \langle V, x\rangle (1 + \frac{\sqrt{d}\zeta}{\|V\|_2}).
\end{aligned}$$

Therefore, we can conclude that

$$\begin{aligned}
\langle \hat{V}, x\rangle - \langle V, x\rangle &\leq \sqrt{d} B_v \|x\|_2 \frac{\delta}{B_v} + \|V\|_2 \|x\|_2 (\frac{\sqrt{d}\zeta}{\|V\|_2}) \\
&\leq (\sqrt{d}\delta + \sqrt{d}\zeta)\|x\|_2 \\
&\leq (\sqrt{d}\delta + \sqrt{d}\zeta) \left((B+\sigma)\sqrt{d} + \sigma \sqrt{2 \ln \frac{B}{\epsilon}} \right).
\end{aligned} \tag{23}$$

Now, for any $\overline{z} \in \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}$, by Lemma 56, we know that we can find a mixture of $m = \lceil \frac{B}{\eta} \rceil^d$ d dimensional Gaussians random variables $\overline{h} = \sum_{i=1}^m w_i g_i$ with bounded means in $[-B, B]^d$ and covariance matrices $\sigma^2 I_d$ such that $d_{TV}(\overline{h}, \overline{z}) \leq 18\sqrt{d}\eta/\sigma$. Let $\overline{\mathcal{H}}$ be the class of all such mixtures.

From Lemma 32, we know that

$$\mathbb{P}\left[\|x\|_2^2 \geq (B+\sigma)\sqrt{d} + \sigma\sqrt{2t} \right] \leq e^{-t}. \tag{24}$$

Setting $t = \ln \frac{B}{\epsilon}$ and $\delta = \zeta = \epsilon/(2dL \ln \frac{B}{\epsilon})$, we can conclude that

$$\mathbb{P}\left[\|x\|_2 \geq B_z \right] = \mathbb{P}\left[\|x\|_2 \geq (B+\sigma)\sqrt{d} + \sigma\sqrt{2 \ln \frac{B}{\epsilon}} \right] \leq \frac{\epsilon}{B}. \tag{25}$$

Therefore, from Equations 23 and 25, we can conclude that for the random variable $\overline{h} = \sum_{i=1}^m w_i g_i$ with $\mathscr{D}(h) = I_h$ and for the coupling $\pi^* \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h}) \right)$ as defined in notations we can write

$$\begin{aligned}
&\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\phi(V^\top \overline{h}) - \phi(\hat{V}^\top \overline{h})\|_2 d\pi^* \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h}) \right) \\
&\leq \int_{Ball_d(0, B_z)} L\sqrt{d}(\delta + \zeta) \left((B+\sigma)\sqrt{d} + \sigma\sqrt{2 \ln \frac{B}{\epsilon}} \right) dI_h \\
&+ \int_{\mathbb{R}^d \setminus Ball_d(0, B_z)} 2B \, dI_h \\
&\leq \frac{(B+\sigma)\epsilon}{2 \ln \frac{B}{\epsilon}} + \frac{\epsilon\sigma}{\sqrt{2d \ln \frac{B}{\epsilon}}} + 2\epsilon,
\end{aligned} \tag{26}$$

where we used the fact that for any $x \in \mathbb{R}^d$, we know that $\|V^\top x - \hat{V}^\top x\|_2$ is bounded and the activation function $\phi(x)$ is Lipschitz continuous with Lipschitz constant L. Here, we assume that the variance of noise is always smaller than 1, i.e., $\sigma \leq 1$. We know that $d \geq 1$ and assuming that $\ln \frac{B}{\epsilon} \geq 1$ (*), we can rewrite Equation 26 as

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\phi(V^\top \overline{h}) - \phi(\hat{V}^\top \overline{h})\|_2 d\pi^* \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h}) \right) \leq (B+1)\epsilon + \epsilon + 2\epsilon \leq (B+4)\epsilon,$$

Then, we have

$$d_{\mathcal{W}}\left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) = \inf_{\pi \in \Pi\left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\phi(V^\top \overline{h}) - \phi(\hat{V}^\top \overline{h})\|_2 d\pi \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right)$$
$$\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\phi(V^\top \overline{h}) - \phi(\hat{V}^\top \overline{h})\|_2 d\pi^* \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) \leq (B+4)\epsilon.$$

Therefore, we have proved that for any $V \in \mathbb{R}^d$ such that $\|V\|_2 \leq B_v$, there exists a vector $\hat{V}$ in C such that for any $\overline{z} \in \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}$ and its estimation with a mixture $\overline{h}$ of Gaussian random variables, we have

$$d_{\mathcal{W}}\left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) \leq (B+4)\epsilon.$$

Case 2 Now, we turn to analyze the case where we have vectors V in $\mathbb{R}^d$ such that $\|V\|_2 > B_v$. We assume that the function ϕ is invertible. Taking into account that ϕ is also bounded in $[-B, B]$, denote $u = \max\left\{|\phi^{-1}(B-\epsilon)|, |\phi^{-1}(-B+\epsilon)|\right\}$. For a given vector $V \in \mathbb{R}^d$, select $b \in A_S$ such that for all $V' \in P_b$, we have $\angle(V, V') \leq \theta$, where θ is defined the same as case 1. From all vectors in P_b, select $\hat{V}$ such that it has the maximum ℓ_2 norm, i.e., the one on the side of the hypercube. It is obvious that $\|\hat{V}\|_2 \geq B_v$. We will show that for any $\overline{h} \in \overline{\mathcal{H}}$, the Wasserstein distance between $\phi(V^\top \overline{h})$ and $\phi(\hat{V}^\top \overline{h})$ is bounded.

Define following two sets

$$\begin{aligned} S_1 &= \{x \in \mathbb{R}^d \mid |\langle V, x\rangle| \leq u\}, \\ S_2 &= \{x \in \mathbb{R}^d \mid |\langle \hat{V}, x\rangle| \leq u\}. \end{aligned} \tag{27}$$

Given any $x \in \mathbb{R}^d \setminus S_1 \cup S_2$ such that $\|x\|_2 \leq B_z$, we show that both of $\langle V, x\rangle$ and $\langle \hat{V}, x\rangle$ are either smaller than $-u$ or larger than u. Assume that $\langle \hat{V}, x\rangle > u$. Denote $\alpha = \angle(\hat{V}, x)$ and $\beta = \angle(V, \hat{V})$. From the fact that $\langle \hat{V}, x\rangle = \|\hat{V}\|_2 \|x\|_2 \cos\alpha \geq u$, we conclude that $\cos\alpha \geq 0$. On the other hand, to conclude that $\langle V, x\rangle$ is also larger than u, we only need to prove that $\langle V, x\rangle \geq 0$ since $x \in \mathbb{R}^d \setminus S_1 \cup S_2$ and we already know that $|\langle V, x\rangle| \geq u$. Therefore, we want to prove that $\langle V, x\rangle = \|V\|_2 \|x\|_2 \cos(\alpha \pm \beta) \geq 0$. It implies that we need to prove $\cos\alpha \geq \sin\beta$. But we know that

$$\begin{aligned} \cos\alpha &\geq \frac{u}{\|\hat{V}\|_2 \|x\|_2} \\ &\geq \frac{u}{\|\hat{V}\|_2 B_z} && \text{(Since } \|x\|_2 \leq B_z\text{)} \\ &\geq \frac{u}{\sqrt{d} B_v B_z} && \text{(Since } \hat{V} \in P_b \text{ and } \|\hat{V}\|_2 \leq \sqrt{d} B_v\text{)} \\ &\geq \frac{B-\epsilon}{L B_v B_z \sqrt{d}} \\ &\geq \frac{\delta}{B_v} \geq \sin\theta \geq \sin\beta, \end{aligned}$$

where we used the fact that the function ϕ is Lipschitz continuous and we know that $|\phi(u) - \phi(-u)| \leq 2Lu$. The last line follows from the fact that $B_z \leq ((B-\epsilon)/\epsilon)\left(2\sqrt{d}\ln(B/\epsilon)\right)$ (**). It is easy to verify in the same way that if $\langle \hat{V}, x\rangle \leq -u$, then $\langle V, x\rangle \leq -u$.

Next, since ϕ is monotone, we can write that for any $x \in \mathbb{R}^d \setminus S_1 \cup S_2$ such that $\|x\|_2 \leq B_z$, we have either both $V^\top x, \hat{V}^\top x$ in $[B-\epsilon, B]$ or both $V^\top x, \hat{V}^\top x$ in $[-B, -B+\epsilon]$, which means that $|V^\top x - \hat{V}^\top x| \leq \epsilon$. Setting $B_v^2 = 2Bu/(\epsilon\sigma\sqrt{2\pi})$, for any mixture of Gaussian random variables

$\overline{h} \in \overline{\mathcal{H}}$ and for the coupling $\pi^* \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right)$, we can write

$$\begin{aligned}
&\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\phi(V^\top \overline{h}) - \phi(\hat{V}^\top \overline{h})\|_2 d\pi^* \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) \\
&\leq \int_{Ball_d(0,B_z)\setminus S_1 \cup S_2} \epsilon d I_h + \int_{S_1 \cup S_2} 2Bd I_h + \int_{\mathbb{R}^d \setminus Ball_d(0,B_z)} 2Bd I_h \\
&\leq \epsilon + 2B \frac{u}{\sqrt{2\pi}\sigma B_v^2} + 2\epsilon \\
&\leq 4\epsilon,
\end{aligned}$$

where we used the fact that $x \in S_1 \cup S_2$ is similar to the probability that $|x| \leq u/B_v$ for the zero mean Gaussian random variable x with variance $(\sigma B_v)^2$. We can, again, write that

$$\begin{aligned}
&d_{\mathcal{W}} \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) = \inf_{\pi \in \Pi\left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\phi(V^\top \overline{h}) - \phi(\hat{V}^\top \overline{h})\|_2 d\pi \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) \\
&\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\phi(V^\top \overline{h}) - \phi(\hat{V}^\top \overline{h})\|_2 d\pi^* \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) \leq 4\epsilon.
\end{aligned}$$

So far, we proved that for any $V \in \mathbb{R}^d$ there exists a $\hat{V} \in C$ such that $d_{\mathcal{W}} \left(\phi(V^\top \overline{h}), \phi(\hat{V}^\top \overline{h})\right) \leq (4+B)\epsilon$ for all mixtures $\overline{h} \in \mathcal{H}$, which comes from the fact that $4\epsilon \leq (4+B)\epsilon$. Now, we turn to covering functions in NET[d,p]. Note that the output of $\phi(V^\top x)$ is real-valued. We also know that Φ is applied element-wise. Consider the set of

$$C_W = \{[V_1^\top \dots V_p^\top]^\top \mid V_i \in C \text{ for } i \in [p]\}.$$

We know that for any $W = [V_1^\top \dots V_p^\top]^\top$ there exists $\hat{W}^\top = [\hat{V}_1^\top \dots \hat{V}_p^\top]^\top$ such that for every $i \in [p]$, we have $d_{\mathcal{W}} \left(\phi(V_i^\top \overline{h}), \phi(\hat{V}_i^\top \overline{h})\right) \leq (4+B)\epsilon$. Therefore, since we keep the coupling the same π^* for every $i \in [p]$, we can conclude that $d_{\mathcal{W}} \left(\Phi(W^\top \overline{h}), \Phi(\hat{W}_i^\top \overline{h})\right) \leq (4+B)\epsilon d$.

Now, using Theorem 20, we get that

$$d_{TV} \left(\overline{g}_\sigma(\Phi(W^\top \overline{h})), \overline{g}_\sigma(\Phi(\hat{W}^\top \overline{h}))\right) \leq \frac{9}{2} \frac{(4+B)\epsilon d}{\sigma} \tag{28}$$

Consequently, for any $\overline{z} \in \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}}_{B,d}$, we can write

$$\begin{aligned}
&d_{TV} \left(\overline{g}_\sigma(\Phi(W^\top \overline{z})), \overline{g}_\sigma(\Phi(\hat{W}^\top \overline{z}))\right) \\
&\leq d_{TV} \left(\overline{g}_\sigma(\Phi(W^\top \overline{z})), \overline{g}_\sigma(\Phi(W^\top \overline{h}))\right) \\
&+ d_{TV} \left(\overline{g}_\sigma(\Phi(W^\top \overline{h})), \overline{g}_\sigma(\Phi(\hat{W}^\top \overline{h}))\right) \\
&+ d_{TV} \left(\overline{g}_\sigma(\Phi(\hat{W}^\top \overline{h})), \overline{g}_\sigma(\Phi(\hat{W}^\top \overline{z}))\right) \\
&\leq \frac{36\sqrt{d}\eta}{\sigma} + \frac{9}{2}(4+B)\frac{\epsilon d}{\sigma},
\end{aligned} \tag{29}$$

where we used data processing inequality and Equation 28. Equation 29 implies that C_W is a global cover for $\overline{\mathcal{G}_\sigma} \circ \text{NET}[d,p]$ with respect to d_{TV} metric. Clearly,

$$|C_W| \leq \left(\frac{(B_v)^{d+1}}{\delta^d \zeta}\right)^p = \left(\frac{2B_v dL \ln \frac{B}{\epsilon}}{\epsilon}\right)^{p(d+1)}.$$

Therefore, setting $\eta = \sqrt{d}(4+B)\epsilon/72$ and $\epsilon' = \epsilon\sigma/(5(4+B)d)$ we conclude that

$$\begin{aligned}
&N_U \left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{NET}[d,p], \infty, d_{TV}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}\right) \leq \left(\frac{10(4+B)d^2 L B_v}{\epsilon\sigma} \ln\left(\frac{5(4+B)Bd}{\epsilon\sigma}\right)\right)^{p(d+1)} \\
&\leq \left(\frac{10\sqrt{10}(4+B)^{3/2}}{(2\pi)^{1/4}} \frac{d^{5/2} L \sqrt{Bu'}}{\epsilon^{3/2}\sigma^2} \ln\left(\frac{5(4+B)Bd}{\epsilon\sigma}\right)\right)^{p(d+1)},
\end{aligned} \tag{30}$$

where

$$u' = \max\left\{|\phi^{-1}(B-\epsilon')|, |\phi^{-1}(-B+\epsilon')|\right\}$$
$$= \max\left\{\left|\phi^{-1}\left(B-\epsilon\sigma/(5(4+B)d)\right)\right|, \left|\phi^{-1}\left(-B+\epsilon\sigma/(5(4+B)d)\right)\right|\right\},$$

and

$$\sigma \leq \frac{5(4+B)Bd}{\epsilon} \leq \frac{5(4+B)Bd}{\epsilon}.$$

Note that we always use $\sigma \leq 1$. In that case, having $\sigma > 5(4+B)Bd/\epsilon$ means that $\epsilon > 5(4+B)Bd > B\sqrt{d}$. On the other hand, the domain of the output of Φ is in $[-B, B]^d$ and, therefore, in this case the covering number would be simply one and no further analysis is required. Furtheremore, the assumotion (*) always hold since in order to obtain an ϵ-cover for the single-layer neural network, we will need to bound the Wassestein distance between $\phi(V^\top \bar{h})$ and $\phi(\hat{V}^\top \bar{h})$ by $(4+B)\epsilon'$. In this case we have

$$\ln \frac{B}{\epsilon'} \geq 1$$
$$\Leftrightarrow \frac{B}{\epsilon'} \geq e$$
$$\Leftrightarrow \frac{B}{e} \geq \frac{\epsilon\sigma}{5(4+B)d}$$
$$\Leftrightarrow \frac{5(4+B)d}{e\sigma} B \geq \epsilon,$$

which holds since we consider $\sigma \leq 1$ and $\epsilon \leq B\sqrt{d}$. Moreover, for assumption (**) to hold, we need

$$B_z \leq \left(\frac{B-\epsilon'}{\epsilon}\right) 2\sqrt{d}\ln(\frac{B}{\epsilon'})$$
$$\Leftrightarrow (B+\sigma)\sqrt{d} + \sigma\sqrt{2\ln\frac{B}{\epsilon'}} \leq \left(\frac{B-\epsilon'}{\epsilon'}\right) 2\sqrt{d}\ln(\frac{B}{\epsilon'})$$
$$\Leftrightarrow \frac{B+1}{\sqrt{\ln\frac{B}{\epsilon'}}} + \frac{\sqrt{2}}{\sqrt{d}} \leq 2\left(\frac{B-\epsilon'}{\epsilon'}\right)\sqrt{\ln\frac{B}{\epsilon'}}$$
$$\Leftrightarrow \frac{B+1}{(\ln\frac{B}{\epsilon'})^{1/4}} + \frac{\sqrt{2}}{\sqrt{d\ln(\frac{B}{\epsilon'})}} \leq 2\left(\frac{B-\epsilon'}{\epsilon'}\right)$$
$$\Leftrightarrow \left(\frac{B+1}{(\ln\frac{B}{\epsilon'})^{1/4}} + \frac{\sqrt{2}}{\sqrt{d\ln(\frac{B}{\epsilon'})}}\right)\frac{\epsilon'}{2} \leq B-\epsilon'$$
$$\Leftrightarrow \left(\frac{B+1+\sqrt{2}}{2}+1\right)\epsilon' \leq B$$
$$\Leftrightarrow \left(\frac{B+3+\sqrt{2}}{2}\right)\left(\frac{\epsilon\sigma}{5(4+B)d}\right) \leq B$$
$$\Leftrightarrow \epsilon \leq \frac{10(4+B)d}{(B+3+\sqrt{2})\sigma} B,$$

which is always true if $\sigma \leq 1$. Note that in both (*) and (**) we were interested in values of ϵ that are smaller than $B\sqrt{d}$; Otherwise, the covering number would be one. □

We can also simplify the constants and write Equation 30 as

$$N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{NET}[d,p], \infty, d_{TV}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,d}}\right) \leq \left(20(4+B)^{3/2}\frac{d^{5/2}L\sqrt{Bu'}}{\epsilon^{3/2}\sigma^2}\ln\left(\frac{5(4+B)Bd}{\epsilon\sigma}\right)\right)^{p(d+1)}.$$

Also since ϕ is a monotone function, we can approximate u' by

$$u' \leq \max\left\{\left|\phi^{-1}\left(B-\frac{\sigma\epsilon}{5(4+B)d}\right)\right|, \left|\phi^{-1}\left(-B+\frac{\sigma\epsilon}{5(4+B)d}\right)\right|\right\}.$$

E Proofs of theorems and and lemmas in Section 7

E.1 Proof of Theorem 26

Proof. We will prove the theorem for the stronger case where the output of single-layer neural network classes and $\mathcal{H}$ is in $[-B, B]^{p_T}$. Consider two consecutive classes NET$[p_i - 1, p_i]$ and NET$[p_i, p_{i+1}]$. From Lemma 18 we know that

$$\begin{aligned}
&N_U\left(\frac{2\epsilon}{2BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}] \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{i-1}, p_i], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_{i-1}}}\right) \\
&\leq N_U\left(\frac{\epsilon}{2BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{i-1}, p_i], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_{i-1}}}\right) \qquad (31) \\
&.N_U\left(\frac{\epsilon}{2BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_i}}\right) = N_i.N_{i+1}.
\end{aligned}$$

Let

$$\overline{\mathcal{Q}} = \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{T-1}, p_T] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_1, p_2].$$

It is clear that $\overline{\mathcal{F}} = \overline{\mathcal{Q}} \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[d, p_1]$. Equation 31 is true for every $2 \leq i \leq T$. Therefore, we can conclude that

$$N_U\left(\frac{(T-1)\epsilon}{2BT\sqrt{p_T}}, \overline{\mathcal{Q}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_1}}\right) \leq \prod_{i=2}^{T} N_i.$$

Using Lemma 18, we can again write that

$$\begin{aligned}
&N_U\left(\frac{\epsilon}{2B\sqrt{p_T}}, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}\right) \\
&\leq N_U\left(\frac{(T-1)\epsilon}{2BT\sqrt{p_T}}, \overline{\mathcal{Q}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_1}}\right).N_U\left(\frac{\epsilon}{2BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[d, p_1], \infty, d_{TV}^{\ell_2}, \overline{\Delta_d}\right) \\
&\leq \prod_{i=1}^{T} N_i.
\end{aligned}$$

Finally, from Theorem 17 and the fact that $\mathcal{F}$ is a class of functions from $\mathbb{R}^d$ to $[-B, B]^p$, we can conclude that

$$N_U\left(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2}\right) \leq N_U\left(\frac{\epsilon}{2B\sqrt{p_T}}, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}\right) \leq \prod_{i=1}^{T} N_i.$$

□

E.2 A technique to build deeper networks from networks with bounded covering number

The following lemma is a technique that can be used to "break" networks in two parts. Then one can find a $\|.\|_2$ covering number for the first few layers and use Theorem 26 for the rest. It is a useful technique that enables the use of existing networks with bounded $\|.\|_2$ covering number to create deeper networks while controlling the capacity. Another possible application of the following lemma is that it gives us the opportunity to get tighter bounds on the covering number in special settings. One example of such settings would be networks that have small norms of weights in the first few layers and potentially large weights in the final layers. In this case, it is possible to use $\|.\|_2$ covering numbers that are dependent on the norms of weights for the first few layers and Theorem 26 for the rest, which does not depend on the norms of weights.

Lemma 35. *Let $\mathcal{Q}$ be a class of functions (e.g., neural networks) from $\mathbb{R}^d$ to $\mathbb{R}^p$ and NET$[p, p_1]$, NET$[p_1, p_2]$, $\ldots$, NET$[p_{T-1}, p_T]$ be T classes of neural networks. Denote the composition of the T-layer neural network and $\mathcal{Q}$ as*

$$\overline{\mathcal{F}} = \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p_{T-1}, p_T] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p_1, p_2] \circ \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p, p_1] \circ \overline{\mathcal{G}_\sigma} \circ \mathcal{Q},$$

and let $\mathcal{H} = \{h : \mathbb{R}^d \to [-B, B]^{p_T} \mid h(x) = \mathbb{E}_{\overline{f}}\left[\,\overline{f}(x)\right], \overline{f} \in \overline{\mathcal{F}}\}$. *Define the uniform covering numbers of composition of neural network classes with the Gaussian noise class (with respect to* d_{TV}^{∞}*) as*

$$N_i = N_U\left(\frac{\epsilon}{4BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p_{i-1}, p_i], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_{i-1}}}\right),\ 1 \le i \le T,\ p_0 = p,$$

and define the uniform covering number of class $\mathcal{Q}$ *as*

$$N_0 = N_U\left(\frac{\epsilon\sigma}{18B\sqrt{p_T}}, \mathcal{Q}, m, \|.\|_2^{\ell_2}\right).$$

Then we have,

$$N_U\left(\epsilon, \mathcal{H}, m, \|.\|_2^{\ell_2}\right) \le \prod_{i=0}^{T} N_i.$$

Proof. From Corollary 21, we can conclude that

$$N_U(\frac{\epsilon}{4B\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \mathcal{Q}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) \le N_U(\frac{\epsilon\sigma}{18B\sqrt{p_T}}, \mathcal{Q}, m, \|.\|_2^{\ell_2}) = N_0.$$

Same as proof of Theorem 26, by using Lemma 18, we can say that for two consecutive classes NET$[p_i - 1, p_i]$ and NET$[p_i, p_{i+1}]$

$$\begin{aligned}
&N_U\left(\frac{2\epsilon}{4BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}] \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{i-1}, p_i], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_{i-1}}}\right) \\
&\le N_U\left(\frac{\epsilon}{4BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{i-1}, p_i], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_{i-1}}}\right) \\
&.N_U\left(\frac{\epsilon}{4BT\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p_i}}\right) = N_i.N_{i+1}
\end{aligned}$$

Let

$$\overline{\mathcal{E}} = \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{T-1}, p_T] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p, p_1].$$

It is clear that $\overline{\mathcal{F}} = \overline{\mathcal{E}} \circ \overline{\mathcal{G}_\sigma} \circ \mathcal{Q}$. Now, from Lemma 18, we can conclude that

$$\begin{aligned}
&N_U\left(\frac{\epsilon}{2B\sqrt{p_T}}, \overline{\mathcal{F}}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}\right) \\
&\le N_U\left(\frac{\epsilon}{4B\sqrt{p_T}}, \overline{\mathcal{E}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{B,p}}\right).N_U(\frac{\epsilon}{4B\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \mathcal{Q}, m, d_{TV}^{\ell_2}, \overline{\Delta_d}) \\
&\le \prod_{i=0}^{T} N_i.
\end{aligned}$$

Lastly, from Theorem 17, we can conclude that

$$N_U(\epsilon, \mathcal{H}, m, \|.\|_2^{\ell_2}) \le N_U\left(\frac{\epsilon}{2B\sqrt{p_T}}, \overline{\mathcal{F}}, \infty, d_{TV}^{\ell_2}, \overline{\Delta_d}\right) \le \prod_{i=0}^{T} N_i.$$

□

F Uniform convergence by bounding the covering number

In this section we provide some technical backgrounds that are related to estimating NVAC and finding valid GBs. Specifically, we discuss how to turn a bound on $\|.\|_2^{\ell_2}$covering number to a bound on generalization gap with respect to ramp loss.

Preliminaries. For any $x \in \mathbb{R}$, the ramp function r_γ with respect to a margin γ is defined as

$$r_\gamma(x) = \begin{cases} 0 & x \le -\gamma, \\ 1 + \frac{x}{\gamma} & [-\gamma, 0], \\ 1 & \gamma > 0. \end{cases}$$

Let $x = [x^{(1)}, \ldots, x^{(k)}]^\top \in \mathbb{R}^k$ be a vector and $\mathcal{Y} = [k]$. The margin function $\mathcal{M} : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ is defined as $\mathcal{M}(x, i) := x^{(i)} - \max_{j \neq i} x^{(j)}$. Next, we define the ramp loss for classification.

Definition 36 (Ramp loss)**.** *Let $f : \mathcal{X} \to \mathbb{R}^k$ be a function and let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = [k]$. We define the ramp loss of function f with respect to margin parameter γ as $l_\gamma(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[r_\gamma(-\mathcal{M}\left(f(x), y\right))\right]$. We also define the empirical counterpart of ramp loss on an input set $S \sim \mathcal{D}^m$ by $\hat{l}_\gamma(f) = \frac{1}{m}\sum_{(x,y)\in S} r_\gamma(-\mathcal{M}(f(x), y))$.*

It is worth mentioning that using (surrogate) ramp loss is a natural case for classification tasks (Bartlett et al., 2006); (Boucheron et al., 2005).

Next, we define the composition of a hypothesis class with the ramp loss function.

Definition 37 (Composition with ramp loss))**.** *Let $\mathcal{F}$ be a hypothesis class from $\mathcal{X}$ to $\mathbb{R}^d$ and $\mathcal{Y} = [k]$. We denote the class of its composition with the ramp loss function by $\mathcal{F}_\gamma : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ and define it as $\mathcal{F}_\gamma = \{(f_\gamma(x, y) = r_\gamma\left(-\mathcal{M}(f(x), y)\right) : f \in \mathcal{F}\}$.*

The following lemma states that we can always bound the 0-1 loss by the ramp loss.

Lemma 38. *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = [k]$ and let f be a function from $\mathcal{X}$ to $\mathbb{R}^k$. We have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l^{0-1}(f(x), y)\right] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[r_\gamma(-\mathcal{M}(f(x), y))\right] = l_\gamma(f).$$

For a proof of Lemma 38 see section A.2 in Bartlett et al. (2017).

One way to bound the generalization gap of a learning algorithm is to find the rate of uniform convergence for class $\mathcal{F}_\gamma$. We define uniform convergence in the following.

Definition 39 (Uniform convergence)**.** *Let $\mathcal{F}$ be a hypothesis class and l be a loss function. We say that $\mathcal{F}$ has uniform convergence property if there exists some function $m_{UC} : (0, 1)^2 \to \mathbb{N}$ such that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and any sample $S \sim \mathcal{D}^m$ if $m \geq m_{UC}(\epsilon, \delta)$ with probability at least $1 - \delta$ (over the randomness of S) for every hypothesis $f \in \mathcal{F}$ we have*

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l(f(x), y)\right] - \frac{1}{m}\sum_{(x,y)\in S} l(f(x), y)\right| \leq \epsilon.$$

An standard approach for finding the rate of uniform convergence is by analyzing the Rademacher complexity of $\mathcal{F}_\gamma$. We now define the empirical Rademacher complexity.

Definition 40 (Empirical Rademacher complexity)**.** *Let $\mathcal{F}$ be a class of hypotheses from $\mathcal{Z}$ to $\mathbb{R}$ and $\mathcal{D}$ be a distribution over $\mathcal{Z}$. The empirical Rademacher complexity of class $\mathcal{F}$ with respect to sample $S = \{z_1, \ldots, z_m\} \sim \mathcal{D}^m$ is denoted by $\hat{\mathfrak{R}}(\mathcal{F}_{|S})$ and is defined as*

$$\hat{\mathfrak{R}}(\mathcal{F}_{|S}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ and σ_i are i.i.d. Rademacher random variables uniformly drawn from $\{0, 1\}$.

The following theorem relates the Rademacher complexity of $\mathcal{F}_\gamma$ to its rate of uniform convergence and provides a generalization bound for the ramp loss and its empirical counterpart on a sample S.

Theorem 41. *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}^k$ and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = [k]$. Let $S \sim \mathcal{D}^m$ denote a sample. Then, for every δ and every $f \in \mathcal{F}$, with probability at least $1 - \delta$ (over the randomness of S) we have*

$$l_\gamma(f) \leq \hat{l}_\gamma(f) + 2\hat{\mathfrak{R}}(\mathcal{F}_{\gamma|S}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

Theorem 41 is an immediate result of standard generalization bounds based on Rademacher complexity (see e.g. Theorem 3.3 in (Mohri et al., 2018)) once we realize that $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f_\gamma\right] = l_\gamma(f)$ and $\frac{1}{m}\sum_{(x,y)\in S} f_\gamma(x, y) = \hat{l}_\gamma(f)$.

We will use Dudley entropy integral (Dudley, 2010) for chaining to bound the Rademacher complexity by covering number; see (Shalev-Shwartz and Ben-David, 2014) for a proof.

Theorem 42 (Dudley entropy integral)**.** *Let $\mathcal{F}$ be a class of hypotheses with bounded output in $[0, c_x]$. Then*

$$\mathfrak{R}(\mathcal{F}_{|S}) \leq \inf_{\epsilon \in [0, c_x/2]} \left\{ 4\epsilon + \frac{12}{\sqrt{m}} \int_{\epsilon}^{c_x/2} \sqrt{\ln N_U(\nu, \mathcal{F}, m, \|.\|_2^{\ell_2})}\, d\nu \right\}.$$

Putting Theorems 41, 42, and Lemma 38 together, we can state the following theorem to bound the 0-1 loss based on the covering number of $\mathcal{F}_\gamma$ and empirical ramp loss.

Theorem 43. *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}^k$ and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y} = [k]$. Let $S \sim \mathcal{D}^m$ be a sample. Then, with probability at least $1 - \delta$ (over the randomness of S) for every $f \in \mathcal{F}$ we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[l^{0-1}(f(x), y) \right] \leq$$
$$l_\gamma(f) \leq \hat{l}_\gamma(f) + \inf_{\epsilon \in [0,1/2]} \left\{ 2 \left[4\epsilon + \frac{12}{\sqrt{m}} \int_{\epsilon}^{1/2} \sqrt{\ln N_U(\nu, \mathcal{F}_\gamma, m, \|.\|_2^{\ell_2})}\, d\nu \right] \right\} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

We will use above theorem in the next appendix to estimate NVAC based on $\|.\|_2^{\ell_2}$ covering number of composition of a class with ramp loss.

G Estimating NVAC using the covering number

In this appendix, we will use Theorem 43 to establish a way of approximating NVAC from a covering number bound. In Remark 44 we state the technique used to approximate NVAC and in the following we will justify why this would be a good approximation.

Remark 44. *Let $\mathcal{F}$ be a hypothesis class from $\mathcal{X}$ to $\mathbb{R}^k$, S be a sample of size m and $\hat{h} \in \mathcal{F}$. We find n^* such that the following holds*

$$\frac{6}{\sqrt{mn^*}} \sqrt{\ln N_U(\epsilon, \mathcal{F}_\gamma, mn^*, \|.\|_2^{\ell_2})} \leq \epsilon, \quad \epsilon = \frac{1 - \hat{l}_\gamma(\hat{h})}{10}, \tag{32}$$

and choose mn^* as an approximation of NVAC. Here, $\hat{l}_\gamma(\hat{h})$ is the empirical ramp loss of $\hat{h}$ on sample S. In Appendix I, where we empirically compare NVAC of different covering number bounds, we choose S to be the MNIST dataset and $\hat{h}$ as the trained neural network (from a class $\mathcal{F}$ of all neural networks with a certain architecture) on this dataset.

In the following we discuss why this choice of mn^* is a good estimate of NVAC. First, let $S^n \in (\mathcal{X} \times \mathcal{Y})^{mn}$ be an input set and $\mathcal{D}$ be a distribution over $(\mathcal{X} \times \mathcal{Y})$, where mn is larger than mn^* as found in Remark 44. From Theorem 43 and using the fact that the ramp loss is in $[0, 1]$ we can write

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[l^{0-1}(\hat{h}(x), y) \right] \leq \hat{l}_\gamma(\hat{h}) + 2\mathfrak{R}(\mathcal{F}_{\gamma_{|S^n}}) + 3\sqrt{\frac{\ln(2/\delta)}{2mn}} \\
&\leq \hat{l}_\gamma(\hat{h}) + \inf_{\epsilon \in [0,1/2]} \left\{ 2 \left[4\epsilon + \frac{12}{\sqrt{mn}} \int_{\epsilon}^{1/2} \sqrt{\ln N_U(\nu, \mathcal{F}_\gamma, mn, \|.\|_2^{\ell_2})}\, d\nu \right] \right\} + 3\sqrt{\frac{\ln(2/\delta)}{2mn}}.
\end{aligned} \tag{33}$$

Since S^n consists of n copies of the sample S, we can replace $\hat{l}_\gamma(\hat{h})$ on S^n by the ramp loss of $\hat{h}$ on S (this would be equal to the ramp loss of trained neural network when we empirically compare NVACs in Appendix I). Moreover, since the number of samples are very large and $\delta = 0.01$, we can approximate the last term in the right hand side of Equation 33 with zero. Therefore, we can write

that

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l^{0-1}(\hat{h}(x),y)\right] \\
&\leq\ \hat{l}_\gamma(\hat{h}) + \inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon + \frac{12}{\sqrt{mn}}\int_\epsilon^{1/2}\sqrt{\ln N_U(\nu,\mathcal{F}_\gamma,mn,\|.\|_2^{\ell_2})}\,d\nu\right]\right\} \\
&\leq \hat{l}_\gamma(\hat{h}) + 2\left[4\epsilon + \frac{12}{\sqrt{mn^*}}\int_\epsilon^{1/2}\sqrt{\ln N_U(\nu,\mathcal{F}_\gamma,mn^*,\|.\|_2^{\ell_2})}\,d\nu\right] \qquad (\forall\epsilon\in[0,1/2]) \\
&\leq\ \hat{l}_\gamma(\hat{h}) + 2\left[4\epsilon + \frac{6}{\sqrt{mn^*}}\sqrt{\ln N_U(\epsilon,\mathcal{F}_\gamma,mn^*,\|.\|_2^{\ell_2})}\right],
\end{aligned}
\tag{34}
$$

where we used the fact that $N_U(\epsilon,\mathcal{F}_\gamma,mn,\|.\|_2^{\ell_2})$ decreases monotonically with ϵ and the integral is over $[\epsilon,1/2]$. Note that in the above equation we subtly used the fact that covering number grows at most polynomially with the number of samples and, therefore, increasing number of samples will always result in smaller right hand side term in Equation 34. In Appendix H, we will show why this is a valid assumption for the covering number bounds that we use in our experiments (see Remark 54).

Since Equation 34 holds for any $\epsilon\in[0,1/2]$, we can set $\epsilon = (1-\hat{l}_\gamma(\hat{h}))/10$ and conclude that

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l^{0-1}(\hat{h}(x),y)\right] \\
&\leq\ \hat{l}_\gamma(\hat{h}) + 2\left[4\epsilon + \frac{6}{\sqrt{mn^*}}\sqrt{\ln N_U(\epsilon,\mathcal{F}_\gamma,mn^*,\|.\|_2^{\ell_2})}\right] \\
&\leq \hat{l}_\gamma(\hat{h}) + 2\,\frac{5(1-\hat{l}_\gamma(\hat{h}))}{10} \\
&\leq 1.
\end{aligned}
\tag{35}
$$

From the above equation, we can conclude that by setting mn to be larger that mn^* as defined in Remark 44, we can provide the following valid generalization bound with respect to l^{0-1} and l_γ:

$$
GB(\hat{h},S^n) = 2\left[4\epsilon + \frac{6}{\sqrt{mn}}\sqrt{\ln N_U(\epsilon,\mathcal{H},mn,\|.\|_2^{\ell_2})}\right].
$$

Moreover, for any S^n such that $mn\geq mn^*$ we can conclude that the GB defined above results in a non-vacuous bound, i.e.,

$$
GB(\hat{h},S^n) + \hat{l}_\gamma(\hat{h}) \leq 1,
$$

which concludes that mn^* is a reasonable approximation for NVAC.

In the next appendix, we discuss different covering number bounds that were mentioned in Section 9. We state these covering number bounds for a general T-layer network in Appendix H. Finally in Appendix I, we present the settings of our experiments and the empirical results of NVAC in Remark 44.

H Different approaches to bound the covering number

In the following, we will state the covering number bounds that were compared in Section 9. We first give two preliminary lemmas. Lemma 46 connects the covering number of a hypothesis class $\mathcal{F}$ to the covering number of $\mathcal{F}_\gamma$, which is used in Remark 44 to obtain generalization bounds. In Lemma 47 we will show a way to find the covering number of a class of functions from $\mathbb{R}^d$ to $\mathbb{R}^p$ from the covering number of real-valued classes that correspond to each dimension. We will use this lemma when we want to compare covering number bounds in the literature that are given for real-valued functions, i.e., Norm-based, Lipschitzness-based, and Pseudo-dim-based approaches.

In the following remark, we will discuss the motivation behind the choice of specific generalization bounds in Section 9

Remark 45 (Choice of generalization bounds)**.** *In our experiments in Section 9 we have not assessed the PAC-Bayes bound in Neyshabur et al. (2018) since it is always looser than the Spectral bound*

of (Bartlett et al., 2017); see Neyshabur et al. (2018) for a discussion. Furthermore, we exclude the generalization bounds that are proved in "two steps". For example, a naive two-step approach is to divide the training data into a large and a small subsets; one can then train the network using the large set and evaluate the resulting hypothesis using the small set. This will give a rather tight generalization bound since in the second step we are evaluating a single hypothesis. However, it does not explain why the learning worked well (i.e., how the learning model came up with a good hypothesis in the first step). More sophisticated two-step approaches such as Dziugaite and Roy (2017); Arora et al. (2018); Zhou et al. (2019) offer more insights on why the model generalizes. However, they do not fully explain why the first step works well (i.e., the prior distribution in Dziugaite and Roy (2017) or the uncompressed network in Arora et al. (2018); Zhou et al. (2019). Therefore, we focus on bounds based on covering numbers (uniform convergence).

Next, we state the preliminaries lemmas that we use in some of the covering number bounds in literature to relate them to covering numbers for the composition of neural networks with the ramp loss.

Lemma 46 (From covering number of $\mathcal{F}$ to covering number of $\mathcal{F}_\gamma$)**.** *Let $\mathcal{F}$ be a hypothesis class of functions from $\mathcal{X}$ to $\mathbb{R}^p$ and $\mathcal{F}_\gamma : \mathcal{X} \times \mathcal{Y} \to [0,1]$ be the class of its composition with ramp loss, where $\mathcal{Y} = [k]$. Then we have*

$$N_U(\epsilon, \mathcal{F}_\gamma, m, \|.\|_2^{\ell_2}) \leq N_U(\frac{\gamma\epsilon}{2}, \mathcal{F}, m, \|.\|_2^{\ell_2}).$$

Proof. First, it is easy to verify that r_γ and $-\mathcal{M}(x,y)$ (with respect to the first input) are Lipschitz continuous functions with respect to $\|.\|_2$ with Lipschitz factors of $1/\gamma$ and 2, respectively; see e.g., section A.2 in Bartlett et al. (2017). Therefore, we can conclude that $r_\gamma\left(-\mathcal{M}(f(x),y)\right)$ is Lipschitz continuous with Lipschitz factor of $2/\gamma$.

Fix an input set $S = \{(x_1,y_1),\ldots,(x_m,y_m)\} \subset \mathcal{X}\times\mathcal{Y}$ and let $C = \{\hat{f}_{i|S} \mid \hat{f}_i \in \mathcal{F},\, i \in [r]\}$ be an $(\gamma\epsilon/2)$-cover for $\mathcal{F}_{|S}$. In the following, we will denote the composition of $\hat{f}_i$ with ramp loss by $\hat{f}_{\gamma,i}$ for the simplicity of notation. Now, we prove that $C_\gamma = \{\hat{f}_{\gamma,i_{|S}} \mid \hat{f}_{\gamma,i} \in \mathcal{F}_\gamma,\, i \in [r]\}$ is also an ϵ-cover for $\mathcal{F}_{\gamma|S}$

Given any $f \in \mathcal{F}$, there exists $\hat{f}_{i|S} \in C$ such that

$$\left\|(\hat{f}_i(x_1),\ldots,\hat{f}_i(x_m)) - (f(x_1),\ldots,f(x_m))\right\|_2^{\ell_2} \leq \frac{\gamma\epsilon}{2}.$$

We can then write that

$$\begin{aligned}
&\left\|(\hat{f}_{\gamma,i}(x_1),\ldots,\hat{f}_{\gamma,i}(x_m)) - (f_\gamma(x_1),\ldots,f_\gamma(x_m))\right\|_2^{\ell_2} \\
&= \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(\hat{f}_{\gamma,i}(x_k) - (f_\gamma(x_k)\right)^2} \\
&\leq \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(r_\gamma\left(-\mathcal{M}(\hat{f}_i(x_k),y_k)\right) - r_\gamma(-\mathcal{M}(f(x_k),y_k))\right)^2}
\end{aligned} \tag{36}$$

From the Lipschitz continuity of $r_\gamma\left(-\mathcal{M}(x,y)\right)$ we can conclude that for any $(x,y) \in \mathcal{X}\times\mathcal{Y}$

$$\left|r_\gamma\left(-\mathcal{M}(f(x),y)\right) - r_\gamma(-\mathcal{M}(\hat{f}_i(x),y))\right| \leq \frac{1}{\gamma}\|\mathcal{M}(\hat{f}_i(x),y) - \mathcal{M}(f(x),y)\|_2 \leq \frac{2}{\gamma}\|\hat{f}_i(x) - f(x)\|_2.$$

Taking the above equation into account, we can rewrite Equation 36 as

$$\begin{aligned}
&\left\|(\hat{f}_{\gamma,i}(x_1),\dots,\hat{f}_{\gamma,i}(x_m)) - (f_\gamma(x_1),\dots,f_\gamma(x_m))\right\|_2^{\ell_2} \\
&\leq \frac{2}{\gamma}\sqrt{\frac{1}{m}\sum_{k=1}^{m}\left((\hat{f}_i(x_k) - f(x_k))\right)^2} \\
&\leq \frac{2}{\gamma}\left\|(\hat{f}_i(x_1),\dots,\hat{f}_i(x_m)) - (f(x_1),\dots,f(x_m))\right\|_2^{\ell_2} \\
&\leq \frac{2}{\gamma}\frac{\gamma\epsilon}{2} \\
&\leq \epsilon.
\end{aligned}$$

In other words, for any $f_{\gamma|_S} \in \mathcal{F}_{\gamma|_S}$ there exists $\hat{f}_{\gamma,i_{|S}} \in S$ such that $\left\|\hat{f}_{\gamma,i_{|S}} - f_{\gamma|_S}\right\|_2^{\ell_2} \leq \epsilon$ and, therefore, C_γ is an ϵ-cover for $\mathcal{F}_{\gamma|_S}$ and the result follows. □

The following lemma finds a covering number for a class of functions with outputs in $\mathbb{R}^p$ from the covering number of the classes of real-valued functions corresponding to each dimension

Lemma 47. *Let $\mathcal{F}_1,\dots,\mathcal{F}_p : \mathcal{X} \to \mathbb{R}$ be p classes of real valued functions. Further let $\mathcal{F} = \left\{f(x) = [f_1(x),\dots,f_p(x)]^\top \mid f_i \in \mathcal{F}_i,\, i \in [p]\right\}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}^p$, where each dimension i in their output comes from the output of a real-valued function in $\mathcal{F}_i$. Then, we have*

$$N_U(\epsilon,\mathcal{F},m,\|.\|_2^{\ell_2}) \leq \prod_{i=1}^{p} N_U(\frac{\epsilon}{\sqrt{p}},\mathcal{F}_i,m,\|.\|_2^{\ell_2}).$$

Proof. Fix an input set $S = \{x_1,\dots,x_m\} \subset \mathcal{X}$. Let $C_1,\dots,C_p$ be $(\epsilon/\sqrt{p})$-covers for $\mathcal{F}_{1|S},\dots,\mathcal{F}_{p|S}$, respectively. We will construct the set C as follows and prove that C is an ϵ-cover for $\mathcal{F}_{|S}$.

$$C = \left\{[\hat{f}_1(x_k),\dots,\hat{f}_p(x_k)]^\top \mid \hat{f}_{i|S} \in C_i,\, i \in [p],\, k \in [m]\right\}.$$

Particularly, from each class $\mathcal{F}_i$, we are choosing all functions $\hat{f}_i$ such that $\hat{f}_{i|S}$ is in C_i. We then use those functions as the dimension i of the output to get functions $f \in \mathcal{F}$. Then we put the restriction of these functions to set S in C. Clearly, $|C| \leq \prod_{i=1}^{p}|C_i|$.

Let $f(x) = [f_1(x),\dots,f_p(x)]^\top$ be any function in $\mathcal{F}$. Since $C_1,\dots,C_p$ are $(\epsilon/\sqrt{p})$-covers for $\mathcal{F}_1,\dots,\mathcal{F}_p$ we know that there exists another set of functions $\hat{f}_i \in \mathcal{F}_i$, $i \in [p]$ such that $\hat{f}_{i|S} \in C_i$ and

$$\left\|(\hat{f}_i(x_1),\dots,\hat{f}_i(x_m)) - (f_i(x_1),\dots,f_i(x_m))\right\|_2^{\ell_2} \leq \frac{\epsilon}{\sqrt{p}}, \quad \forall i \in [p].$$

Let $\hat{f}(x) = [\hat{f}_1(x), \ldots, \hat{f}_p(x)]^\top$. We can then write that

$$\begin{aligned}
\left\| f_{|S} - \hat{f}_{|S} \right\|_2^{\ell_2} &= \left\| (f(x_1), \ldots, f(x_m)) - (\hat{f}(x_1), \ldots, \hat{f}(x_m)) \right\|_2^{\ell_2} \\
&= \sqrt{\frac{1}{m} \sum_{k=1}^{m} \left\| f(x_k) - \hat{f}(x_k) \right\|_2^2} \\
&\leq \sqrt{\frac{1}{m} \sum_{k=1}^{m} \sum_{i=1}^{p} \left(f_i(x_k) - \hat{f}_i(x_k) \right)^2} \\
&\leq \sqrt{\sum_{i=1}^{p} \sum_{k=1}^{m} \frac{1}{m} \left(f_i(x_k) - \hat{f}_i(x_k) \right)^2} \\
&\leq \sqrt{\sum_{i=1}^{p} \left(\left\| (f_i(x_1), \ldots, f_i(x_m)) - (\hat{f}_i(x_1), \ldots \hat{f}_i(x_m)) \right\|_2^{\ell_2} \right)^2} \\
&\leq \sqrt{\sum_{i=1}^{p} \frac{\epsilon^2}{p}} \\
&\leq \epsilon
\end{aligned}$$

Therefore, we can conclude that C is an ϵ-cover for $\mathcal{F}_{|S}$. Since $|C| \leq \prod_{i=1}^{p} |C_i|$ the result follows. □

In the following we will state the covering number bounds that are compared using their NVACs in Section 9 .

Covering number bounds. We first state the bound in Theorem 26, where we use the covering number of Theorem 25 for each layer of the neural network.

Theorem 48 (The bound of Theorem 26)**.** *Let NET[d, p_1], NET[p_1, p_2], $\ldots$, NET[p_{T-1}, p_T] be T classes of neural networks. Denote the T-layer noisy network by*

$$\overline{\mathcal{F}} = \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p_{T-1}, p_T] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p_1, p_2] \circ \overline{\mathcal{G}_\sigma} \circ \textit{NET}[d, p_1],$$

and let $\mathcal{H} = \{h : \mathbb{R}^d \to [0,1]^{p_T} \mid h(x) = \mathbb{E}_{\overline{f}} \left[\overline{f}(x) \right], \overline{f} \in \overline{\mathcal{F}}\}$. Then we have

$$\begin{aligned}
&\ln N_U \left(\epsilon, \mathcal{H}_\gamma, m, \|.\|_2^{\ell_2} \right) \\
&\leq \sum_{i=2}^{T} p_i . p_{i-1} \ln \left(282 \frac{(4T\sqrt{p_T})^{3/2} p_{i-1}^{5/2} \sqrt{\ln \left(\frac{(120/\gamma) T \sqrt{p_T} p_{i-1} - \epsilon\sigma}{\epsilon\sigma} \right)}}{(\gamma\epsilon)^{3/2}\sigma^2} \ln \left(\frac{120 T p_{i-1} \sqrt{p_T}}{\gamma\epsilon\sigma} \right) \right) \\
&+ d p_1 \ln \left(\frac{18 T e m \sqrt{p_T}}{\gamma\epsilon\sigma} \right).
\end{aligned}$$

Proof. We first use Theorem 25 to find the covering number of NET[p_{i-1}, p_i]. Particularly, for any $2 \leq i \leq T$ we have,

$$\begin{aligned}
&\ln N_i = \ln N_U \left(\frac{\epsilon}{2T\sqrt{p_T}}, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{i-1}, p_i], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{1,p_{i-1}}} \right) \\
&\leq p_i . p_{i-1} \ln \left(282 \frac{(2T\sqrt{p_T})^{3/2} p_{i-1}^{5/2} \sqrt{\ln \left(\frac{60 T \sqrt{p_T} p_{i-1} - \epsilon\sigma}{\epsilon\sigma} \right)}}{\epsilon^{3/2}\sigma^2} \ln \left(\frac{60 T p_{i-1} \sqrt{p_T}}{\epsilon\sigma} \right) \right).
\end{aligned}$$

Moreover, we use Lemma 14.17 in Anthony et al. (1999) to find a bound on N_0. This lemma provides a bound with respect to $\|.\|_2^\infty$, however, we know that $\|.\|_2^{\ell_2}$ is always smaller than $\|.\|_2^\infty$ (see Remark 5). Therefore, we can bound N_0 as follows

$$\ln N_0 \leq dp_1 \ln\left(\frac{9Tem\sqrt{p_T}}{\epsilon\sigma}\right).$$

From Theorem 26 we know that $\ln N_U\left(\epsilon, \mathcal{H}, m, \|.\|_2^{\ell_2}\right) \leq \sum_{i=1}^{T} \ln N_i$, therefore, we can write that

$$\begin{aligned}
&\ln N_U\left(\epsilon, \mathcal{H}, m, \|.\|_2^{\ell_2}\right) \\
&\leq \sum_{i=1}^{T} p_i.p_{i-1} \ln\left(282\frac{(2T\sqrt{p_T})^{3/2}p_{i-1}^{5/2}\sqrt{\ln\left(\frac{60T\sqrt{p_T}p_{i-1}-\epsilon\sigma}{\epsilon\sigma}\right)}}{\epsilon^{3/2}\sigma^2}\ln\left(\frac{60Tp_{i-1}\sqrt{p_T}}{\epsilon\sigma}\right)\right) \\
&+ dp_1 \ln\left(\frac{9Tem\sqrt{p_T}}{\epsilon\sigma}\right).
\end{aligned}$$

Applying Lemma 46 to turn this covering number into a covering number for $\mathcal{H}_\gamma$ concludes the result. □

Notation. For the rest of this section, we will be presenting the covering number bounds in literature. For some of these bounds we will require the norm of the weights of network. Therefore, we use a new notation for the class of single-layer neural networks. More precisely, we will use NET$[d, p, W]$ to denote the same class as NET$[d, p]$ but we are also pointing out the weight matrix $W \in \mathbb{R}^{d\times p}$. For a matrix $W \in \mathbb{R}^{d\times p}$ we denote its $\|.\|_{s,t}$ norm as $\|(\|W_{:,1}\|_s, \ldots, \|W_{:,p}\|_s)\|_t$, where $W_{:,i}$ denotes the ith column of W (e.g. for a weight matrix W, $\|W^\top\|_{1,\infty}$ refers to the maximum of $\|.\|_1$ norm of incoming weights of a neuron). By $\|W\|_\sigma$ we denote the spectral norm of a matrix. For a matrix $X \in \mathbb{R}^{d\times m}$ we denote its normalized Frobenious norm by$\|X\|_F$, which is defined as $\|X\|_F = \sqrt{\frac{1}{m}\sum x_{i,j}^2}$.

We would like to mention that, in the experiments, we use a slightly different form of sigmoid function for the activation function rather than the one in Definition 23. Indeed, we will add a constant to the sigmoid function to turn it into an odd function in $[-1/2, 1/2]$. In the following remark we will discuss the reason behind this choice and the fact that it does not change the covering number in Theorem 48.

Remark 49. *The bound in the Spectral covering number requires the activation functions to output 0 at the origin. Therefore, in our experiments in Section 9, we set $\phi(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$ as activation functions for neurons of the network, so that $\phi(0) = 0$ and $\phi(x) \in [-1/2, 1/2]$. This will not affect the covering number bound of Theorem 48. The bound in Theorem 48 is derived from the covering number bound of Theorem 25 for single-layer neural network classes. There are three sources of dependency on the activation function in Theorem 25. The first one is the dependence on the range of output, which is 1 for both $\phi(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$ and the sigmoid function ($\phi(x) = \frac{1}{1+e^{-x}}$) defined in Definition 23. The second dependecy is the Lipschitz factor which is 1 for both of the activation functions. The final dependency is on $u = \max\left\{\left|\phi^{-1}(B-\epsilon)\right|, \left|\phi^{-1}(-B+\epsilon)\right|\right\}$. It is easy to verify that the value of u for $\phi(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$ is exactly the same as the value of u for $\phi(x) = \frac{1}{1+e^{-x}}$. As a result, using both $\phi(x) = \frac{1}{1+e^{-x}}$ and $\phi(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$ will result in the same covering number bound in Theorem 48. Generally, adding a constant to the output of functions in a class will not change its covering number.*

We will now discuss the Norm-based bound from Theorem 14.17 in Anthony et al. (1999), which is a bound for real-valued networks. Therefore, we will apply Lemma 47 to relate it to a covering number for neural networks with p output dimensions.

Theorem 50 (Norm-based covering number)**.** *Let NET$[d, p_1, W_1]$, $\ldots$, NET$[p_{T-1}, p_T, W_T]$ be T classes of neural networks and $\mathcal{F}$ = NET$[p_{T-1}, p_T, W_T] \circ \ldots \circ$ NET$[d, p_1, W_1]$. Denote by V the*

maximum of $\|.\|_{1,\infty}$ among the layers of the network, i.e., $V = \max_i \|W_i^\top\|_{1,\infty}$. Then we have

$$\log_2 N_U(\epsilon, \mathcal{F}_\gamma, m, \|.\|_2^{\ell_2}) \leq \frac{\sqrt{p_T}}{2}(\frac{2\sqrt{p_T}}{\gamma\epsilon})^{2T}(2V)^{T(T+1)}\log_2(2d+2).$$

Proof. The proof simply follows from Theorem 14.17 in Anthony et al. (1999) and Lemmas 46 and 47 once we realize that the sigmoid function is Lipschitz continuous with Lipschitz factor of 1. □

Next we state the Pseudo-dim-based bound.

Theorem 51 (Psuedo-dim-based covering number)**.** *Let NET[d, p_1], ..., NET[p_{T-1}, p_T] be T classes of neural networks and $\mathcal{F}$ = NET[p_{T-1}, p_T] $\circ \ldots \circ$ NET[d, p_1]. Denote the total number of weights of network by $W = dp_1 + \sum_{i=2}^{T} p_{i-1}.p_i$ and the total number of hidden neurons by $r = \sum_{i=1}^{T} p_i$. Furthermore, let P be as follows*

$$P = ((W+2)r)^2 + 11(W+2)r\log_2\left(18(W+2)r^2\right).$$

Then we have

$$\ln N_U(\epsilon, \mathcal{F}_\gamma, m, \|.\|_2^{\ell_2}) \leq \sqrt{p_T}P\ln\left(\frac{2\sqrt{p_T}em}{P\gamma\epsilon}\right).$$

Proof. By Theorem 14.2 in Anthony et al. (1999) we know that the pseudo dimension (P_{dim}) of $\mathcal{F}_i$ is smaller or equal to P, where $\mathcal{F}_i$ is the class of real-valued functions corresponding ith dimension of output of functions in class $\mathcal{F}$ (for a definition of pseudo dimension see for instance Chapter 11 in Anthony et al. (1999)). Furthermore, from the standard analysis of covering number and pseudo dimension (see e.g., Theorem 12.2 in Anthony et al. (1999)), we can write

$$\ln N_U(\epsilon, \mathcal{F}_i, m, \|.\|_2^{\ell_2}) \leq P_{\text{dim}}\ln(\frac{em}{\epsilon P_{\text{dim}}}).$$

Combining the above equation with Lemmas 46 and 47 concludes the result. □

Now we turn into presenting the Lipschitzness-based bound.

Theorem 52 (Lipschitzness-based covering number)**.** *Let NET[d, p_1, W_1], ..., NET[p_{T-1}, p_T, W_T] be T classes of neural networks and $\mathcal{F}$ = NET[p_{T-1}, p_T, W_T] $\circ \ldots \circ$ NET[d, p_1, W_1]. Denote by V the maximum of $\|.\|_{1,\infty}$ among all but the first layers of the network, i.e., $V = \max_{i\geq 2} \|W_i^\top\|_{1,\infty}$ and denote the total number of weights of network by $W = dp_1 + \sum_{i=2}^{T} p_{i-1}.p_i$. Then we have*

$$\ln N_U(\epsilon, \mathcal{F}_\gamma, m, \|.\|_2^{\ell_2}) \leq W\sqrt{p_T}\ln\left(\frac{4em\sqrt{p_T}WV^T}{\gamma\epsilon(V-1)}\right).$$

Proof. The covering number follows from the bound in Theorem 14.5 in Anthony et al. (1999), which is a $\|.\|_2^\infty$ covering number, but we know that $\|.\|_2^{\ell_2}$ is always smaller than $\|.\|_2^\infty$. Therefore, from Theorem 14.5 in Anthony et al. (1999), Lemma 47, and the fact that sigmoid is a Lipschitz continuous function with Lipschitz factor of 1 we know that

$$\ln N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2}) \leq W\sqrt{p_T}\ln\left(\frac{2em\sqrt{p_T}WV^T}{\epsilon(V-1)}\right).$$

Combining the above equation with Lemma 46 will result in the desired bound. □

Finally, we will present the Spectral bound in Bartlett et al. (2017). In our experiments, we consider the reference matrices M_i in the following theorem to be zero.

Theorem 53 (Spectral covering number)**.** *Let NET[d, p_1, W_1], ..., NET[p_{T-1}, p_T, W_T] be T classes of neural networks and $\mathcal{F}$ = NET[p_{T-1}, p_T, W_T] $\circ \ldots \circ$ NET[d, p_1, W_1]. Let reference matrices $M_1 \in \mathbb{R}^{d\times p_1}$ and $M_i \in \mathbb{R}^{p_{i-1}\times p_i}$, $2 \leq i \leq T$ be given. For an input set $S = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ define $X = [x_1 \ldots x_m] \in \mathbb{R}^{d\times m}$ as the collection of input samples. Further, for any $i \in [T]$, let $\|W_i\|_\sigma \leq s_i$ and $\|W_i^\top - M_i^\top\|_{2,1} \leq b_i$. Then we have*

$$\ln N_U(\epsilon, \mathcal{F}_\gamma, m, \|.\|_2^{\ell_2}) \leq \frac{4\|X\|_F^2\ln(2W^2)}{\gamma^2\epsilon^2}\left(\prod_{i=1}^{T} s_i^2\right)\left(\sum_{i=1}^{T}\left(\frac{b_i}{s_i}\right)^{2/3}\right)^3.$$

The original bound in Bartlett et al. (2017) considers the input norm $\|X\|_F^2$ to be the sum of $\|.\|_2^2$ norms of input samples and adjusts the chaining technique of Theorem 42 to account for this assumption. Here, for the sake of consistency, we consider the Forbenious norm to be normalized and use the conventional chaining technique, which applies to the $\|.\|_2^{\ell_2}$ metric.

Remark 54. *Some of the bounds that we presented are dependent on the number of input samples m. However, for all of them the logarithm of covering number has at most a logarithmic dependence on the number of samples. It is also worth mentioning that the Spectral bound is dependent on the normalized Frobenious norm and increasing the number of copies of S in Equation 34 (i.e., mn)will not change this norm and, therefore, the Spectral bound.*

I Empirical results

In this appendix we will discuss details of the learning settings for the empirical results that were stated in Section 9. We train fully connected neural networks on the publicly available MNIST dataset, which consists of handwritten digits (28×28 pixel images) with 10 labels. Our baseline architecture has 3 hidden layers each containing 250 neurons, one input layer, and one output layer. The input layer has 784 neurons, which are pixels of each image in MNIST dataset. The output layer has 10 neurons, corresponding to the 10 labels. All the activation functions are the shifted variant of the sigmoid function as discussed in Appendix H, i.e., $\phi(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$. The additional architecures that we use are as follows: (a) fully connected neural networks with one input layer, one output layer, and $2, 4, 5$ hidden layers each containing 250 neurons; (b) fully connected neural networks with one input layer, one output layer, and three hidden layers each containing $64, 150, 350, 500, 800, 1000, 1500$ neurons. All of the experiments are performed using NVIDIA Titan V GPU.

Networks are trained with SGD optimizer with a momentum of 0.9 and a learning rate of 0.3. For the purpose of training the loss is set to be the cross-entropy loss. For the rest of the experiments (e.g., to report the accuracy and NVACs) ramp loss with a margin of $\gamma = 0.1$ is used. The size of training, validation, and test sets are 59000, 1000, and 10000, respectively. In Theorem 26 we are considering noisy networks with its expectation as output. Therefore, for reporting results of Theorem 26 we compute the output 50 times and take an average. Computing random outputs several times and averaging them yields in negligible error bars in the demonstrated results.

The results of NVAC as a function of depth and width are depicted in Figure 2. All of the NVACs are derived according to Remark 44. In Figure 2, we also include the Norm-based approach (Theorem 50) which was omitted from the Figures in Section 9 due to its large scale. As mentioned in Section 9, the bounds that are based on norms and group norms perform poorly compared to those that are based on parameter count. In the following, we will investigate this observation.

The first justification behind this observation is the dependence on $1/\epsilon$. From Theorems 50 and 53 we know that the logarithm of covering number in Norm-based approach has a polynomial dependence on $1/\epsilon$, i.e., $O((1/\epsilon)^{2T})$, while in the Spectral approach it has a linear dependence, i.e, $O(1/\epsilon)$. On the other hand, Pseudo-dim-based, Lipschitzness-based, and Theorem 26 has a logarithmic dependece on $1/\epsilon$.

The second reason behind this observation is that the Spectral and Norm-based approaches depend on the product of the weights. Although one may think that in networks with large number of parameters this dependency would be better than those on the number of parameters, we will see that the Pseudo-dim-based, Lipschitzness-based, and Theorem 26 perform better in these cases. For instance, consider the netwrok that has been trained with three hidden layers, each containing 1500 neurons. In this case, the number of parameters is $\approx 5 \times 10^9$, while in the Spectral approach, the contribution of product of norms to covering number is $\approx 1 \times 10^9$ and the contribution of $1/\epsilon$ is $\approx 4 \times 10^4$. In the norm-based approach the contribution of the product of norms is $\approx 1 \times 10^{53}$ alone.

Finally, we will explore the observation that the bound in Theorem 26 performs better than the Pseudo-dim-based and Lipschitzness-based bounds. Let w denote the maximum number of neurons in a hidden layer. The logarithm of the covering number bound in Theorem 26 depends on $O(w^2 \ln(w^{5/2}))$, while the Pseudo-dim-based bound in Theorem 51 depends on $O(w^6)$. Comparing Theorem 26 with the Lipschitzness-based covering number is more challenging because Lipschitzness-based bound depends on $O(w^2 \ln(w^2))$ and also on $T \ln(V)$. It is also important to note that Theorem 26 works naturally for multi-output layer while Lipschitzness-based bound works

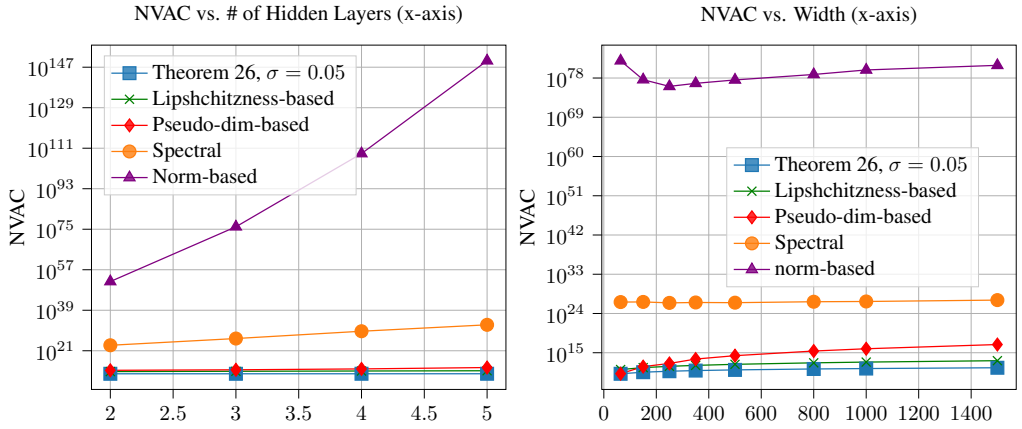

Figure 2: NVAC of different generalization bounds as a function of the number of hidden layers and width of the network.

for real-valued functions and requires to find a covering number for each output separately. The empirical results, however, suggests that the Lipschitzness-based bound is worse than the bound in Theorem 26. It is worth mentioning that in the rightmost graph in Figure 1, the output of noisy networks are averaged over 1000 noisy outputs to obtain results that are more close to the true expectation that has been considered in the output of architecture in Theorem 26.

J Techniques to estimate smooth densities with mixtures of Gaussians

Notation. Denote by $\mathscr{D}(\overline{x})$ the probability density function of the random variable $\overline{x}$. Let $1\{x \in S\}$ be an indicator function that outputs 1 if $x \in S$ and 0 if $x \notin S$. For a function $f : \mathcal{X} \to \mathcal{Y}$, let $f_+(x) = \max\{0, f(x)\}$ and $f_-(x) = \min\{0, f(x)\}$. By $\mathbb{R}^d \setminus [-B, B]^d$ we refer to the complement of set $[-B, B]^d$ with respect to $\mathbb{R}^d$. We also denote by $f * g$ the convolution of functions f and g. For two sets S_1 and S_2, we define their Cartesian product by $S_1 \times S_2$ and by S^d we refer to the Cartesian power, i.e., $S^d = \{(s_1, \ldots, s_d) \mid s_i \in S, \forall i \in [d]\}$. In the following lemma, we sometimes drop the overlines in our notation and simply write x when we are referring to random variables. When it is clear from the context, we write f instead of $f(x)$.

Lemma 55 (Gaussian kernel estimation of bounded distributions)**.** *Let $\overline{x}$ be a random variable in $\overline{\mathcal{X}_{B,d}}$ and denote its probability density function by $f = \mathscr{D}(\overline{x})$. Let g be the density function of a zero mean Gaussian random variable with covariance matrix $\sigma^2 I_d$. Given a set $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ of i.i.d. samples $x_i \sim f$, $i \in [n]$, we define the empirical measure as $\mu_n(x) = \frac{1\{x \in S\}}{n}$. Then, we have*

$$\mathbb{E}\left[\int_{\mathbb{R}^d} |(\mu_n * g)(x) - (f * g)(x)|\, dx\right] \leq 2\sqrt{\frac{1}{n}}\left(\frac{2B}{\sqrt{(2\pi\sigma^2)}} + 1\right)^d$$

Proof. Note that $\int \mu_n(x)dx = 1$ and since f and g are probability density functions, we know that $\int (f * g)(x)dx = 1$ and $\int (\mu_n * g)(x)dx = 1$. Therefore, we have (for simplicity, we write $\mathbb{E}_{x_i \sim f}$

instead of $\mathbb{E}_{\substack{x_i \sim f, \\ i \in [n]}}$)

$$
\begin{aligned}
&\mathbb{E}_{x_i \sim f}\left[\int_{\mathbb{R}^d} |(\mu_n * g)(x) - (f * g)(x)|\, dx\right] \\
&= \int_{\mathbb{R}^d} \mathbb{E}_{x_i \sim f}\left[|(\mu_n * g)(x) - (f * g)(x)|\, dx\right] \\
&= 2\int_{\mathbb{R}^d} \mathbb{E}_{x_i \sim f}\left[(\mu_n * g - f * g)_+ (x) dx\right] \\
&\leq 2\int_{\mathbb{R}^d} \sqrt{\mathbb{E}_{x_i \sim f}\left[((\mu_n * g)(x) - (f * g)(x))^2\right]} dx \qquad \text{(By Jensen's inequality)} \\
&\leq 2\int_{\mathbb{R}^d} \sqrt{\mathbb{E}_{x_i \sim f}\left[\left(\frac{1}{n}\sum_{i=1}^n g(x - x_i) - \int f(y) g(x - y) dy\right)^2\right]} dx.
\end{aligned} \tag{37}
$$

Now, we can write

$$
\begin{aligned}
&\mathbb{E}_{x_i \sim f}\left[\left(\frac{1}{n}\sum_{i=1}^n g(x - x_i) - \int f(y) g(x - y) dy\right)^2\right] = \mathbb{E}_{x_i \sim f}\left[\left(\frac{1}{n}\sum_{i=1}^n g(x - x_i)\right)^2\right] \\
&+ \mathbb{E}_{x_i \sim f}\left[\left(\int f(y) g(x - y) dy\right)^2\right] - \mathbb{E}_{x_i \sim f}\left[2\left(\frac{1}{n}\sum_{i=1}^n g(x - x_i)\right)\left(\int f(y) g(x - y) dy\right)\right] \\
&= \mathbb{E}_{x_i \sim f}\left[\left(\frac{1}{n}\sum_{i=1}^n g(x - x_i)\right)^2\right] + \left(\int f(y) g(x - y) dy\right)^2 \\
&- 2\left(\int f(y) g(x - y) dy\right)\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x_i \sim f}\left[g(x - x_i)\right]\right) \\
&= \mathbb{E}_{x_i \sim f}\left[\left(\frac{1}{n}\sum_{i=1}^n g(x - x_i)\right)^2\right] - \left(\int f(y) g(x - y) dy\right)^2,
\end{aligned} \tag{38}
$$

where the last equality comes from the fact that the expectation is over random variables $x_1, \ldots, x_n$

$$
\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x_i \sim f}\left[g(x - x_i)\right] = \frac{1}{n}\sum_{i=1}^n \int g(x - y) f(y) dy = \int g(x - y) f(y) dy = f * g.
$$

Next, we know that

$$
\begin{aligned}
&\mathbb{E}_{x_i \sim f}\left[\left(\frac{1}{n}\sum_{i=1}^n g(x - x_i)\right)^2\right] = \frac{1}{n^2}\mathbb{E}_{x_i \sim f}\left[\left(\sum_{i=1}^n g(x - x_i)\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}_{x_i \sim f}\left[\sum_{i=1}^n g(x - x_i)^2\right] + \frac{1}{n^2}\mathbb{E}\left[\sum_{i \neq j}^n g(x - x_i) g(x - x_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{x_i \sim f}\left[g(x - x_i)^2\right] + \frac{1}{n^2}\sum_{i \neq j}^n \mathbb{E}_{x_i, x_j \sim f}\left[g(x - x_i) g(x - x_j)\right] \\
&= \frac{1}{n}\mathbb{E}_{x_i \sim f}\left[g(x - x_i)^2\right] + \frac{1}{n^2}\sum_{i \neq j}^n \mathbb{E}_{x_i \sim f}\left[g(x - x_i)\right]\mathbb{E}_{x_j \sim f}\left[g(x - x_j)\right] \\
&= \frac{1}{n}\mathbb{E}_{x_i \sim f}\left[g(x - x_i)^2\right] + (1 - \frac{1}{n})\left(\mathbb{E}_{x_i \sim f}\left[g(x - x_i)\right]\right)^2 \\
&= \frac{1}{n}\mathbb{E}_{x_i \sim f}\left[g(x - x_i)^2\right] + (1 - \frac{1}{n})\left(\int g(x - y) f(y) dy\right)^2.
\end{aligned} \tag{39}
$$

Putting Equations 39 and 38 together, we have

$$\begin{aligned}
&\mathbb{E}_{x_i\sim f}\left[\left(\frac{1}{n}\sum_{i=1}^{n} g(x-x_i) - \int f(y)g(x-y)dy\right)^2\right] \\
&= \frac{1}{n}\mathbb{E}_{x_i\sim f}\left[g(x-x_i)^2\right] - \frac{1}{n}\left(\int g(x-y)f(y)dy\right)^2 \\
&= \frac{1}{n}\int g(x-y)^2 f(y)dy - \frac{1}{n}\left(\int g(x-y)f(y)dy\right)^2 \\
&= \frac{1}{n}\left(f*g^2 - (f*g)^2\right).
\end{aligned} \tag{40}$$

Therefore, we can rewrite Equation 37 as

$$\begin{aligned}
&\mathbb{E}_{x_i\sim f}\left[\int_{\mathbb{R}^d} |(\mu_n * g)(x) - (f*g)(x)|\, dx\right] \\
&\leq 2\int_{\mathbb{R}^d} \sqrt{\frac{1}{n}(f*g^2 - (f*g)^2)}dx \\
&\leq 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d} \sqrt{(f*g^2 - (f*g)^2)}dx.
\end{aligned} \tag{41}$$

We know that g is the probability density function of $\mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. Consequently, we know that

$$g(x)^2 = \frac{1}{(2\pi)^d\sigma^{2d}}\exp(-\frac{1}{\sigma^2}x^\top x) \leq \frac{1}{(2\pi\sigma^2)^d},$$

and we can rewrite Equation 41 as

$$\begin{aligned}
&\mathbb{E}_{x_i\sim f}\left[\int_{\mathbb{R}^d} |(\mu_n * g)(x) - (f*g)(x)|\, dx\right] \\
&\leq 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d} \sqrt{(f*g^2 - (f*g)^2)}dx \leq 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d} \sqrt{f*g^2}dx \\
&\leq 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d} \sqrt{\int g(x-y)^2 f(y)dy}\, dx \\
&= 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d} \sqrt{\int \frac{1}{(2\pi\sigma^2)^d}\exp\left(-\frac{1}{\sigma^2}(x-y)^\top(x-y)\right) f(y)dy}\, dx \\
&= 2\sqrt{\frac{1}{n}}\int_{[-B,B]^d} \sqrt{\int \frac{1}{(2\pi\sigma^2)^d}\exp\left(-\frac{1}{\sigma^2}(x-y)^\top(x-y)\right) f(y)dy}\, dx \\
&+ 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d\setminus[-B,B]^d} \sqrt{\int \frac{1}{(2\pi\sigma^2)^d}\exp\left(-\frac{1}{\sigma^2}(x-y)^\top(x-y)\right) f(y)dy}\, dx \\
&\leq 2\sqrt{\frac{1}{n}}\int_{[-B,B]^d} \sqrt{\int \frac{1}{(2\pi\sigma^2)^d} f(y)dy}\, dx \\
&+ 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d\setminus[-B,B]^d} \sqrt{\frac{1}{(2\pi\sigma^2)^d}\int \exp\left(-\frac{1}{\sigma^2}(x-y)^\top(x-y)\right) f(y)dy}\, dx.
\end{aligned}$$

We can then conclude that

$$
\begin{aligned}
&\mathbb{E}_{x_i\sim f}\left[\int_{\mathbb{R}^d}|(\mu_n * g)(x) - (f*g)(x)|\,dx\right]\\
&\le 2\sqrt{\frac{1}{n}}\int_{[-B,B]^d}\sqrt{\frac{1}{(2\pi\sigma^2)^d}}\,dx\\
&+ 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d\setminus[-B,B]^d}\sqrt{\frac{1}{(2\pi\sigma^2)^d}\int\exp\left(-\frac{1}{\sigma^2}(x-y)^\top(x-y)\right)f(y)dy}\,dx\\
&\le 2\sqrt{\frac{1}{n}}\frac{(2B)^d}{\sqrt{(2\pi\sigma^2)^d}} + 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d\setminus[-B,B]^d}\sqrt{\frac{1}{(2\pi\sigma^2)^d}\int\exp\left(-\frac{1}{\sigma^2}(x-y)^\top(x-y)\right)dy}\,dx\\
&\le 2\sqrt{\frac{1}{n}}\frac{(2B)^d}{\sqrt{(2\pi\sigma^2)^d}} + 2\sqrt{\frac{1}{n}}\int_{\mathbb{R}^d\setminus[-B,B]^d}\sqrt{\frac{1}{(2\pi\sigma^2)^d}}\int\exp\left(-\frac{1}{2\sigma^2}(x-y)^\top(x-y)\right)dy\,dx\\
&\le 2\sqrt{\frac{1}{n}}\frac{(2B)^d}{\sqrt{(2\pi\sigma^2)^d}} + 2\sqrt{\frac{1}{n}}\sum_{i=1}^{d}\binom{d}{i}\frac{(2B)^{d-i}\sqrt{(2\pi)^i}\sigma^i}{\sqrt{(2\pi\sigma^2)^d}}\\
&\le 2\sqrt{\frac{1}{n}}\sum_{i=0}^{d}\binom{d}{i}\frac{(2B)^{d-i}}{\sqrt{(2\pi\sigma^2)^{d-i}}} = 2\sqrt{\frac{1}{n}}\left(\frac{2B}{\sqrt{(2\pi\sigma^2)}}+1\right)^d.
\end{aligned}
\tag{42}
$$

Here, we used the fact that for f is supported on $[-B,B]^d$ and the maximum value of $\exp(-(1/\sigma^2)(x-y)^\top(x-y))$ is 1 over $[-B,B]^d$. Moreover, for a fixed x in $\mathbb{R}^d\setminus[-B,B]^d$, the maximum value of $\exp(-(1/\sigma^2)(x-y)^\top(x-y))$ happens when $(x-y)^\top(x-y)$ is minimized, therefore, Whenever $x^{(i)} > B$, the minimization occurs when $y^{(i)} = B$. On the other hand, when $x^{(i)} < B$, the minimization happens when $y^{(i)} = -B$. We can, then, consider the integration over $\mathbb{R}^d\setminus[-B,B]^d$ as sum of integrals over subsets where for some $i\in[d], |x^{(i)}| > B$. Then we can upper bound the integration over each subset by the marginalization of the Gaussian variable in dimensions where $|x^{(i)}| > B$ and consider the fact that the exponent is always smaller than the exponent of an i dimensional Gaussian distribution in those subsets. Note that, when we use this lemma, we consider large values of n such that the expectation of our kernel estimation can get as small as desired. It is also noteworthy that the upper bound on the expectation implies that there exists a set of samples $S = \{x_1,\dots,x_n\}$ that can achieve the desired upper bound. □

Lemma 55 can be used to estimate any bounded distribution that is perturbed with Gaussian noise with a mixture of Gaussians with bounded means and equal diagonal covariance matrix. To do so, we can first use Lemma 55 to approximate the distributions with Gaussian kernels over n i.i.d samples from the distribution. We can then divide the subset $[-B,B]^d$ into several subsets and define a Gaussian on each subset that has a weight equal to the number of samples on each interval. We provide the formal version of this estimation in the following lemma.

Lemma 56. *Let $\overline{x}\in\overline{\mathcal{X}_{B,d}}$ be a random variable and denote its probability density function by $f = \mathscr{D}(\overline{x})$. Let g be the density function of a zero mean Gaussian random variable with covariance matrix $\sigma^2 I_d$. Then for any small value η, we can estimate $f*g$ by a mixture of $\lceil\frac{B}{\eta}\rceil^d$ Gaussians $\sum_{i=1}^k g(x-\mu_i)$, where $\mu_i\in[-B,B]^d$ and*

$$d_{TV}(f*g, \sum_{i=1}^{k} g(x-\mu_i)) \le \frac{18\sqrt{d}\eta}{\sigma}$$

Proof. From Lemma 55, we know that there exists a set $S=\{x_1,\dots,x_n\}\subset\mathbb{R}^d$ of i.i.d. samples from f and its empirical measure $\mu_n(x) = \frac{1\{x\in S\}}{n}$ such that the total variation between f and the sum of Gaussian kernels defined on empirical measure is bounded

$$d_{TV}\left(f*g, \sum_{i=1}^{n} g(x-x_i)\right) \le 2\sqrt{\frac{1}{n}}\left(\frac{2B}{\sqrt{(2\pi\sigma^2)}}+1\right)^d = \epsilon.$$

Denote $m = \lceil \frac{B}{\eta} \rceil$. We construct the following grid P of points on $[-B, B]^d$ and choose means of the Gaussian densities based on it

$$P = \{-B + 2i\eta \mid i \in [m]\}^d.$$

For any $a = (a_1, \ldots, a_d) \in [m]^d$, we define

$$\mu_a = \left[-B + (2a_1 + 1)\eta, \ \ldots \ , -B + (2a_d + 1)\eta\right]^\top \in \mathbb{R}^d$$

as a choice of mean vector for the Gaussian mixture. We claim that by choosing appropriate weights, we can estimate $f * g$ with respect to total variation distance by a mixture of Gaussians with means in the following set

$$M = \left\{\mu_a = [\mu_a^{(1)} \ldots \mu_a^{(d)}]^\top \in \mathbb{R}^d \mid \mu_a^{(i)} = -B + (2a_i + 1)\eta, \ \forall a = (a_1, \ldots, a_d) \in [m]^d\right\}.$$

For the set $S = \{x_1, \ldots, x_n\}$ that was sampled for kernel estimate $\mu_n * g$, we choose the weight w_a for the Gaussian density with mean μ_a as follows. Define the set S_a as

$$S_a = \{x_i \in S \mid x_i \in [-B + 2a_1\eta, -B + 2(a_1 + 1)\eta] \times \ldots \times [-B + 2a_d\eta, -B + 2(a_d + 1)\eta]\} \tag{43}$$

Next, we select w_a as

$$w_a = \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \in S_a\} = \frac{|S_a|}{n}.$$

In other words, w_a is the number of samples in S that the ℓ_∞ distance between those samples and μ_a is smaller than 2η. Note that the cardinality of M, which is the number of Gaussian densities in the mixture is $|M| = (\lceil \frac{B}{\eta} \rceil)^d$.

We now prove that the total variation distance between $\mu_n * g$ and $\sum_{a\in[m]^d} w_a g(x - \mu_a)$ is smaller than $\frac{9\sqrt{d}}{\sigma}\eta$.

$$\begin{aligned}
&d_{TV}\left(\frac{1}{n}\sum_{i=1}^{n} g(x - x_i), \sum_{a\in[m]^d} w_a g(x - \mu_a)\right) \\
&= \frac{1}{2}\left\|\frac{1}{n}\sum_{i=1}^{n} g(x - x_i) - \sum_{a\in[m]^d} w_a g(x - \mu_a)\right\|_1 \\
&= \frac{1}{2}\left\|\sum_{a\in[m]^d}\left(\frac{1}{n}\sum_{x_i\in S_a} g(x - x_i) - w_a g(x - \mu_a)\right)\right\|_1 \\
&\leq \frac{1}{2}\sum_{a\in[m]^d}\left\|\left(\frac{1}{n}\sum_{x_i\in S_a} g(x - x_i) - w_a g(x - \mu_a)\right)\right\|_1 \quad \text{(By triangle inequality).}
\end{aligned} \tag{44}$$

Now, we can write

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{x_i\in S_a} g(x - x_i) - w_a g(x - \mu_a)\right\|_1 \\
&\leq \left\|\frac{1}{n}\sum_{x_i\in S_a} (g(x - x_i) - g(x - \mu_a))\right\|_1 \quad \text{(Since } w_a = \frac{|S_a|}{n}\text{)} \\
&\leq \frac{1}{n}\sum_{x_i\in S_a} \|g(x - x_i) - g(x - \mu_a)\|_1 .
\end{aligned} \tag{45}$$

From Theorem 30, we know that

$$\begin{aligned}
&2d_{TV}\left(g(x - x_i) - g(x - \mu_a)\right) = \|g(x - x_i) - g(x - \mu_a)\|_1 \\
&\leq 9\frac{\|x_i - \mu_a\|_2}{\sigma} \leq \frac{9\sqrt{d}}{\sigma} 2\eta.
\end{aligned} \tag{46}$$

Putting Equation 46 into Equation 45, we have

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{x_i\in S_a} g(x-x_i) - w_a g(x-\mu_a)\right\|_1 \\ &\leq \frac{1}{n}\sum_{x_i\in S_a}\frac{9\sqrt{d}}{\sigma}2\eta = w_a\frac{9\sqrt{d}}{\sigma}2\eta.\end{aligned} \tag{47}$$

Now, putting Equations 45 and 47 together, we can rewrite Equation 44 as

$$\begin{aligned}&d_{TV}\left(\frac{1}{n}\sum_{i=1}^{n} g(x-x_i), \sum_{a\in[m]^d} w_a g(x-\mu_a)\right) \\ &\leq \frac{1}{2}\sum_{a\in[m]^d}\left\|\left(\frac{1}{n}\sum_{x_i\in S_a} g(x-x_i) - w_a g(x-\mu_a)\right)\right\|_1 \\ &\leq \frac{1}{2}\sum_{a\in[m]^d} w_a\frac{9\sqrt{d}}{\sigma}2\eta \\ &= \frac{9\sqrt{d}}{\sigma}\eta.\end{aligned} \tag{48}$$

Note that the bound in Equation 48 does not depend on the size of sampled set S. Therefore, we can choose n as large as we want. Specifically, we choose n as follows

$$n = \left(\frac{2B}{\sqrt{2\pi\sigma^2}} + 1\right)^{2d} \cdot \left(\frac{9\sqrt{d}}{2\sigma}\eta\right)^{-2}$$

We can then conclude that for any random variable $\overline{x}$ defined over $[-B, B]^d$, we can approximate the density function of $\overline{x} + \overline{z}, \overline{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$ with a mixture of $\lceil\frac{B}{\eta}\rceil^d$ Gaussians with means in $[-B, B]^d$ such that

$$d_{TV}\left(f * g, \sum_{a\in[m]^d} w_a g(x-\mu_a)\right) \leq \epsilon + \frac{9\sqrt{d}}{\sigma}\eta = \frac{18\sqrt{d}\eta}{\sigma}.$$

□