ANNUAL REVIEWS

Annual Review of Statistics and Its Application

A Survey on Statistical Theory of Deep Learning: Approximation, Training Dynamics, and Generative Models

Namjoon Suh and Guang Cheng

Department of Statistics & Data Science, University of California, Los Angeles, California, USA; email: guangcheng@stat.ucla.edu

Annu. Rev. Stat. Appl. 2025. 12:177–207

First published as a Review in Advance on November 21, 2024

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040522-013920

Keywords

deep learning, approximation theory, training dynamics, generative adversarial networks, diffusion model, in-context learning

Abstract

In this article, we review the literature on statistical theories of neural networks from three perspectives: approximation, training dynamics, and generative models. In the first part, results on excess risks for neural networks are reviewed in the nonparametric framework of regression. These results rely on explicit constructions of neural networks, leading to fast convergence rates of excess risks. Nonetheless, their underlying analysis only applies to the global minimizer in the highly nonconvex landscape of deep neural networks. This motivates us to review the training dynamics of neural networks in the second part. Specifically, we review articles that attempt to answer the question of how a neural network trained via gradient-based methods finds a solution that can generalize well on unseen data. In particular, two well-known paradigms are reviewed: the neural tangent kernel and mean-field paradigms. Last, we review the most recent theoretical advancements in generative models, including generative adversarial networks, diffusion models, and in-context learning in large language models from two of the same perspectives, approximation and training dynamics.

1. INTRODUCTION

In recent years, the field of deep learning (Goodfellow et al. 2016) has experienced a substantial evolution. Its impact has transcended traditional boundaries, leading to significant advancements in sectors such as health care (Esteva et al. 2019), finance (Heaton et al. 2017), autonomous systems (Grigorescu et al. 2020), and natural language processing (Otter et al. 2020). Neural networks, the mathematical abstractions of our brain, lie at the core of this progression. Nevertheless, amid the ongoing renaissance of artificial intelligence (AI), neural networks have acquired an almost mythical status, spreading the misconception that they are more art than science. It is important to dispel this notion. While the applications of neural networks may evoke awe, they are firmly rooted in mathematical principles. In this context, the importance of deep learning theory becomes evident. Several key points underscore its significance.

1.1. Why Is Theory Important?

In this subsection, we aim to emphasize the importance of understanding deep learning within mathematical and statistical frameworks. Here are some key points to consider:

1. Deep learning is a dynamic and rapidly evolving field, producing hundreds of thousands of publications online. Today's models are characterized by highly intricate network architectures comprising numerous complex subcomponents (e.g., Transformer; Vaswani et al. 2017). Amidst this complexity, it becomes crucial to comprehend the fundamental principles underlying these models, and placing these models within a unified mathematical framework is essential. Such a framework serves as a valuable tool for distilling the core concepts from these intricate models, making it possible to extract and comprehend the key principles that drive their functionality.
2. Applying statistical frameworks to deep learning models allows meaningful comparisons with other statistical methods. For instance, widely used statistical estimators like wavelet or kernel methods can prompt questions about when and why deep neural networks might perform better. This analysis helps us understand when deep learning excels compared with traditional statistical approaches, benefiting both theory and practice.
3. Hyperparameters, such as learning rate, weight initializations, network architecture choices, activation functions, etc., significantly influence the quality of the estimated model. Understanding the proper ranges for these hyperparameters is essential not only for theorists but also for practitioners. For instance, in the era of big data, where there are millions of samples in one dataset, the theoretical wisdom tells us the depth of the network should scale logarithmically in sample size for the good estimation of compositional functions (see, e.g., Schmidt-Hieber 2020).

In this review, we provide an overview of articles that delve into these concepts in mathematical settings, offering readers specific insights into the topics discussed above. Here, we try to avoid too many technicalities and make the introductions accessible to as many statisticians as possible. Some more technical components can be found in the **Supplemental Appendix**.

Supplemental Material >

1.2. Road Map of the Article

We classify the existing literature on statistical theories of neural networks into three categories, discussed in Sections 2–4, respectively.

1. Approximation theory: Recently, much work has been done to bridge the approximation theory of neural network models (Hornik et al. 1989, Mhaskar 1996, Yarotsky 2017,

Petersen & Voigtlaender 2018, Hanin 2019, Montanelli & Du 2019, Schmidt-Hieber 2020, Blanchard & Bennouna 2022) and the tools in empirical processes (Van de Geer 2000) to obtain the fast convergence rates of excess risks in both regression (Schmidt-Hieber 2020, Hu et al. 2021) and classification tasks (T. Hu et al. 2020, Kim et al. 2021) under nonparametric settings. Approximation theory provides useful perspectives in measuring the fundamental complexities of neural networks for approximating functions in certain classes. Specifically, it enables the explicit construction of neural networks for the function approximations so that we know how the network width, depth, and number of active parameters should scale in terms of sample size, data dimension, and the function smoothness index to get good statistical convergence rates. For simplicity, we mainly consider the works in which the fully connected neural networks are used as the function estimators. These works include those of Schmidt-Hieber (2020), Kim et al. (2021), Shen et al. (2021), Jiao et al. (2021), Lu et al. (2021), Imaizumi & Fukumizu (2019, 2022), Suzuki (2018), Suzuki & Nitanda (2021), Chen et al. (2022a), and Suh et al. (2022) under various problem settings. Yet, these works assume that the global minimizers of loss functions are obtainable, and they are mainly interested in the statistical properties of these minimizers without any optimization concerns. However, this is a strong assumption, given the nonconvexity of loss functions arising from the nonlinearities of activation functions in the hidden layers.

2. Training dynamics: Understanding the landscape of nonconvex loss functions for neural network models and its impact on their generalization capabilities represents a critical next step in the literature. However, the nontrivial nonconvexity of this landscape poses significant challenges for the mathematical analysis of many intriguing phenomena observed in neural networks. For example, the seminal empirical finding of Zhang et al. (2021) reveals that neural networks in their experiments trained on a standard image classification training set (CIFAR-10) can fit the (noisy) training data perfectly and, at the same time, show respectable prediction performance (see Zhang et al. 2021, figure 1*c*). This contradicts the classic statistical wisdom of the bias–variance trade-off, which states that overfitted models cannot generalize well. The role of overparametrizations (e.g., Bartlett et al. 2021) on the nonconvex optimization landscape of neural networks has been intensively studied over the past few years, and we review the relevant literature in this context. For instance, Jacot et al. (2018) revealed that the dynamics of highly overparametrized neural networks with large enough width, trained via gradient descent (GD) in ℓ_2-loss, behave similarly to those of functions in reproducing kernel Hilbert spaces (RKHSs), where the kernel is associated with a specific network architecture. Many subsequent works study the training dynamics and the generalization abilities of neural networks in the kernel regime under various settings (Nitanda & Suzuki 2020, Hu et al. 2021). However, due to technical constraints (as detailed in Section 3.1), networks in the kernel regime fail to explain the essential functionality of neural networks—feature learning (Zhong et al. 2016). Another important line of work focuses on understanding the learning dynamics of neural networks in the mean-field (MF) regime, where feature learning becomes more explainable. Nonetheless, the analysis in the MF regime is challenging to generalize to deep networks and requires infinite widths. Finally, we conclude this section by presenting several approaches that go beyond or unify the two regimes.
3. Generative modeling: In this section, we review the most recent theoretical advancements in generative models, including generative adversarial networks (GANs), diffusion models, and in-context learning (ICL) in large language models (LLMs). The works introduced are based on the philosophies of two paradigms (approximation and training dynamics). Over the past decade, GANs (Goodfellow et al. 2014) have stood out as a significant unsupervised

learning approach, known for their ability to learn the data distributions and efficiently sample the data from them. In this review, we discuss articles that study the statistical properties of GANs (Arora et al. 2017, Zhang et al. 2018, Bai et al. 2019, Liang 2021, Schreuder et al. 2021, Chen et al. 2022b). Recently, another set of generative models, diffusion models, have shown superior performance to GAN models in generating impressive qualities of synthetic data in various data modalities, including image data (Song et al. 2020, Dhariwal & Nichol 2021), tabular data (Kim et al. 2022, Suh et al. 2023), and medical imaging (Müller-Franzes et al. 2022). However, given diffusion models' complex nature and recent introduction to the community, the theoretical reasons why they work so well remain vague. Lastly, we review the interesting phenomenon of ICL, which is commonly observed in LLMs. This refers to the ability of LLMs conditioned on a prompt sequence consisting of examples from a task (input–output pairs) along with the new query input to generate the corresponding output accurately. Readers can refer to the nice survey articles of Gui et al. (2021) and Yang et al. (2024) for detailed descriptions of the methodologies and applications of GANs and diffusion models in various domains. Dong et al. (2024) provide an overview of ICL that highlights some key findings and advancements.

In relation to Section 1.1, the advantages of neural networks over classic statistical function estimators are primarily discussed in Sections 2 and 3 under various problem settings. In Section 3, we review the work of Yang & Hu (2022), which suggests appropriate parameter initialization scalings and learning rates for feature learning in large-scale (infinite width) neural networks.

1.3. Existing Surveys on Deep Learning Theory

To our knowledge, there are four existing review articles on deep learning theory (Bartlett et al. 2021, Belkin 2021, Fan et al. 2021, He & Tao 2021). There is overlap in certain subjects covered by each of the articles, but their main focuses differ. Bartlett et al. (2021) provided a comprehensive and technical survey of statistical understandings of deep neural networks. In particular, they focused on examining the significant influence of overparametrization in neural networks, which plays a key role in enabling gradient-based methods to discover interpolating solutions. Fan et al. (2021) introduced the most commonly employed neural network architectures in practice, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), and training techniques such as batch normalization and dropout. They also provided a brief introduction to the approximation theory of neural networks. Like Bartlett et al. (2021), Belkin (2021) reviewed the role of overparametrization for implicit regularization and benign overfitting, observed not only in neural network models but also in classic statistical models, such as weighted nearest neighbor predictors. Most notably, they provided intuitions on the roles of the overparametrization of nonconvex loss landscapes of neural networks through the lens of optimization. He & Tao (2021) provided a comprehensive overview of deep learning theory, including the ethics and security problems that arise in data science and their relationships with deep learning theory. We recommend readers review all these articles to gain a comprehensive understanding of this emerging field. Our article offers a unique and comprehensive survey of the statistical results of neural networks, focusing on approximation theory and training dynamics, while also covering generative models within these two paradigms.

2. APPROXIMATION THEORY–BASED STATISTICAL GUARANTEES

We outline fully connected networks, which are the main object of interest throughout this review. From a high-level perspective, deep neural networks can be viewed as a family of nonlinear

statistical models that can encode highly nontrivial representations of data. The specific network architecture $(L, \mathbf{p})$ consists of a positive integer L, called the number of hidden layers, and a width vector $\mathbf{p} := (\mathbf{p}_0, \ldots, \mathbf{p}_{L+1}) \in \mathbb{N}^{L+2}$, recording the number of nodes from input to output layers in the network. A fully connected neural network, $\tilde{f}$, is then any function of the form for the input feature $\mathbf{x} \in \mathcal{X}$:

$$\tilde{f} : \mathcal{X} \to \mathbb{R}, \quad \mathbf{x} \to f(\mathbf{x}) = W_L \sigma W_{L-1} \sigma W_{L-2} \ldots \sigma W_1 \mathbf{x}, \qquad 1.$$

where $\mathbf{W}_i \in \mathbb{R}^{p_{i+1} \times p_i}$ is a weight matrix with $\mathbf{p}_0 = d$, $\mathbf{p}_{L+1} = 1$ and $\boldsymbol{\sigma}$ is the nonlinear activation function. Here, the activation function plays a key role in the neural network allowing the nonlinear representations of the given data $\mathbf{x}$. Popular examples include rectified linear units (ReLU), $\mathrm{ReLU}(x) = \max(x, 0)$, and $\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}}$. We omit the bias terms added on the outputs of preactivated hidden layers for simplicity, but bias terms are needed for universal approximation if the input data are not appended with a constant entry.

Under this setting, complexity of the networks is mainly measured through the three metrics: (*a*) the maximum width, denoted as $\mathbf{p}_{\max} := \max_{i=0,\ldots,L+1} \mathbf{p}_i$; (*b*) the depth, denoted as L; and (*c*) the number of nonzero parameters, denoted as $\mathcal{N}$. Letting $\|\mathbf{W}_j\|_0$ be the number of nonzero entries of $\mathbf{W}_j$ in the jth hidden layer, the final form of the neural network we consider is given by

$$\mathcal{F}(L, \mathbf{p}, \mathcal{N}) := \left\{ \tilde{f} \text{ of the form in Equation 1} : \sum_{j=1}^{L} \|\mathbf{W}_j\|_0 \le \mathcal{N} \right\}. \qquad 2.$$

In the approximation theoretic literature, the capacity or expressive power of a neural network is often characterized by the tuple $(L, \mathbf{p}_{\max}, \mathcal{N})$. Let $\mathcal{G}$ be a function class where the target function $f_\star$ belongs. The main question frequently asked is, given the fixed approximation error, ε, defined as

$$\varepsilon := \sup_{f_\star \in \mathcal{G}} \inf_{f \in \mathcal{F}(L, \mathbf{p}, \mathcal{N})} \|f - f_\star\|_{L_p}, \qquad 3.$$

how does the network architecture $(L, \mathbf{p}_{\max}, \mathcal{N})$ scale in terms of ε? Note the supremum is taken over the function class $\mathcal{G}$ and the infimum is taken over the neural network class $\mathcal{F}$. The distance between two functions is measured via L_p norm.

2.1. Expressive Power of Fully Connected Networks

In this section, we briefly review some important results in the approximation theory of neural networks. For a more comprehensive review, readers can refer to DeVore et al. (2021).

2.1.1. Approximating functions in $\mathcal{G}$. The specifications of function classes $\mathcal{G}$ and $\mathcal{F}$ allow us to derive many interesting insights on the power of neural networks. For instance, the celebrated universal approximation theorem states that any continuous functions (i.e., $\mathcal{G} := \{\text{continuous functions on } \mathbb{R}^d\}$) can be approximated by a shallow neural network (i.e., one hidden layer) with a sigmoid activation function (i.e., $\mathcal{F} := \{\text{shallow neural networks}\}$) at an arbitrary accuracy (Cybenko 1989; Hornik et al. 1989, 1990). However, achieving a good approximation may require an extremely large number of hidden nodes, which significantly increases the capacity of $\mathcal{F}$. Barron (1993, 1994) developed an approximation theory for function classes $\mathcal{G}$ with limited capacity, measured by the integrability of their Fourier transform. Interestingly, the approximation result is not affected by the dimension of input data d, and this observation matches with the experimental results that deep learning is very effective in dealing with high-dimensional data.

Nonetheless, the capacity of $\mathcal{G}$ in the work of Barron (1994) is rather limited. Another typical route of the analysis is to specify the smoothness of function classes. Roughly, smoothness

refers to the highest order of derivatives that the functions can possess. Notably, Yarotsky (2017) demonstrated that deep ReLU networks (Equation 1) cannot escape the curse of dimensionality when approximating functions in the unit ball in Sobolev space. Yarotsky (2017) established that the order $\mathcal{N} = \mathcal{O}(\varepsilon^{-\frac{d}{r}})$ is sharp, with matching lower and upper bounds. Petersen & Voigtlaender (2018) generalized the results to the class of piecewise smooth functions. Later, Schmidt-Hieber (2020) developed a theory that for any network architecture satisfying the set of conditions on $(L, \mathbf{p}_{\max}, \mathcal{N})$, deep ReLU nets can achieve good approximation rates for functions in Hölder classes. (More details on the technical results of Schmidt-Hieber (2020) are provided in **Supplemental Appendix A**.) This should be contrasted with the result of Yarotsky (2017) that proved the existence of a network with good approximation. Many researchers have been working on considering either more general (i.e., Besov space) or more specific (i.e., hierarchical compositional function) function classes $\mathcal{G}$ than Hölder classes. These considerations have facilitated numerous intriguing comparisons between classic statistical function estimators and deep neural networks in terms of their fundamental limits, specifying the second item in Section 1.1.

Supplemental Material >

2.1.2. The benefits of depth. Several studies have shown that the expressive power of deep neural networks grows with respect to the number of layers (L). Delalleau & Bengio (2011) showed there exist families of functions that can be represented much more efficiently with a deep network than with a shallow one (i.e., one with substantially fewer hidden units). In the asymptotic limit of depth, Pascanu et al. (2014) showed deep ReLU networks can represent exponentially more piecewise linear functions than their single-layer counterparts, given that both networks have the same number of nodes. Montufar et al. (2014) proved that a similar result can be derived with the fixed number of hidden layers. Poole et al. (2016) showed deep neural networks can disentangle highly curved manifolds in an input space into flat manifolds in a hidden space, while shallow networks cannot. Mhaskar et al. (2017) demonstrated that deep ReLU networks can approximate compositional functions with significantly fewer parameters (i.e., $\mathcal{N}$) than shallow neural networks need in order to achieve the same level of approximation accuracy.

2.1.3. Bounded width. The effects of width on the expressive power of neural networks have recently been studied (Lu et al. 2017, Hanin 2019, Kidger & Lyons 2020, Park et al. 2020, Vardi et al. 2022). Lu et al. (2017) showed that the minimal width for universal approximation (denoted as $w_{\min}$) using ReLU networks with respect to the L_1 norm of functions from $\mathbb{R}^d \to \mathbb{R}$ is $d+1 \le w_{\min} \le d+4$. Kidger & Lyons (2020) extended the results to L_p-approximation of functions from $\mathbb{R}^d \to \mathbb{R}^{\text{out}}$ and obtained $w_{\min} \le d + d_{\text{out}} + 1$. Park et al. (2020) further improved $w_{\min} = \max\{d+1, d_{\text{out}}\}$. Universal approximations of narrow networks with other activation functions were studied by Park et al. (2020), Kidger & Lyons (2020), and Johnson (2018). Note that the aforementioned works require the depth of networks to be exponential in input dimension with bounded width, which are the dual versions of the universal approximation of bounded depth networks from Cybenko (1989) and Hornik et al. (1989, 1990). Interestingly, Vardi et al. (2022) provided evidence that the width of the networks can be less important than depth for the expressive power of neural nets. They showed that the price for making the width small is only a linear increase in the network depth, in sharp contrast with the results mentioned earlier on how making the width small may require an exponential increase in the network depth.

2.2. Statistical Guarantees for Regression Tasks

The natural question is, What are the interpretations or consequences of the results in approximation theory for deriving statistical guarantees of neural networks under noisy observations? In this section, we focus on reviewing several important results under regression tasks in this regard. First, we introduce the settings frequently adopted in statistical learning theory.

Let $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ be the measurable feature space and output space. We denote as ρ a joint probability measure on the product space $\mathcal{X} \times \mathcal{Y}$ and let $\rho_{\mathcal{X}}$ be the marginal distribution of the feature space $\mathcal{X}$. We assume that the noisy dataset $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ is generated from the nonparametric regression model,

$$\mathbf{y}_i = f_\rho(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{4.}$$

where the noise ε_i is assumed to be a centered random variable and $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0$. Our goal is to estimate the regression function $f_\rho(\mathbf{x})$ with the given noisy dataset $\mathcal{D}$. Here, it is easy to see that the regression function $f_\rho := \mathbb{E}(\mathbf{y}|\mathbf{x})$ is a minimizer of the population risk $\mathcal{E}(f)$ under ℓ_2-loss defined as

$$\mathcal{E}(f) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\Big[\big(\mathbf{y} - f(\mathbf{x})\big)^2\Big].$$

However, since the joint distribution ρ is unknown, we cannot find f_ρ directly. Instead, we solve the following empirical risk minimization problem induced from the dataset $\mathcal{D}$:

$$\widehat{f}_n = \underset{f\in\mathcal{F}(L,\mathbf{p},\mathcal{N})}{\arg\min}\ \mathcal{E}_D(f) := \underset{f\in\mathcal{F}(L,\mathbf{p},\mathcal{N})}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^n \big(\mathbf{y}_i - f(\mathbf{x}_i)\big)^2\right\}. \tag{5.}$$

Note that the articles referenced in this section always assume the empirical risk minimizer $\widehat{f}_n$ is obtainable, ignoring the optimization process. The function estimator f is structurally regularized by $\mathcal{N}$ in $(L, \mathbf{p}, \mathcal{N})$, which is specified below.

Under this setting, the excess risk is an important statistical object measuring the generalizability of the function estimator $\widehat{f}_n$ for unseen data in $\mathcal{X}$. Mathematically, it can be shown that it is a difference between population risks of f_ρ and $\widehat{f}_n$ (see Wainwright 2019, chapter 13), which is $\mathbb{E}_{\mathbf{X}\sim\rho_{\mathcal{X}}}\big[(\widehat{f}_n(\mathbf{X}) - f_\rho(\mathbf{X}))^2\big]$. The excess risk can be further decomposed as follows (Suh et al. 2022, proposition 4.2):

$$\mathbb{E}_{\mathbf{X}\sim\rho_{\mathcal{X}}}\big[(\widehat{f}_n(\mathbf{X}) - f_\rho(\mathbf{X}))^2\big] \le \frac{\text{Complexity measure of } \mathcal{F}}{n} + (\text{Approximation error})^2. \tag{6.}$$

In the context of excess risk bounds, it is important to note the trade-off between the approximation error and the combinatorial complexity measure of a neural network class $\mathcal{F}$. Specifically, as the network hypothesis space $\mathcal{F}$ becomes richer, the approximation results improve. However, increasing the hypothesis space $\mathcal{F}$ arbitrarily will eventually lead to an increase in the complexity measure of $\mathcal{F}$, as described in Equation 6. Researchers (e.g., Bartlett et al. 2019, Schmidt-Hieber 2020) have examined how various complexity measures, including Vapnik–Chervonenkis dimension (VC-dimension), pseudodimension, and covering number, scale with respect to $(L, \mathbf{p}_{\max}, \mathcal{N})$. Specifically, these papers proved all three complexity measures increase linearly in $\mathcal{N}$. For achieving good convergence rates of the excess risks from Equation 6, it is crucial to properly specify the network architecture [i.e., the choices of $(L, \mathbf{p}_{\max}, \mathcal{N})$] that balances the tension between the complexity of $\mathcal{F}$ and approximation error in terms of sample size n, data dimension d, and function smoothness $r \ge 0$.

2.2.1. Deep sparse ReLU networks versus linear estimators. Among the list of articles to be discussed, the seminal work of Schmidt-Hieber (2020), which first appeared on arXiv in 2017, paved the way for providing the statistical guarantees of deep ReLU networks in the sense of Equation 6. Schmidt-Hieber (2020) demonstrated that sparsely connected deep ReLU networks (Equation 2) significantly outperform traditional statistical estimators. Specifically, if the unknown regression function f_ρ is a composition of functions that are individually estimable faster than $\mathcal{O}\big(n^{-\frac{2r}{2r+d}}\big)$, then a composition-based deep ReLU network is provably more effective than estimators that do not utilize compositions, such as wavelet estimators. For further discussion, readers

are directed to Kutyniok (2020), Ghorbani et al. (2020), Shamir (2020), and Kohler & Langer (2020).

The sparse network structure manifested in $\mathcal{N}$ in the article had already been proven to be impressively effective in the compressed learning literature (Iandola et al. 2016; Han et al. 2015, 2016). The sparsity of networks can be achieved via pruning technique (Han et al. 2015). Iandola et al. (2016) empirically showed that the pruned CNNs with 50 times fewer parameters achieve the same accuracy level as AlexNet (Krizhevsky et al. 2012) in image classification tasks, and these results pave the way for the employment of neural networks in small devices such as smart phones or smart watches.

In the statistical literature, after the publication of Schmidt-Hieber (2020), several other works analyzed sparsely connected networks. Imaizumi & Fukumizu (2019) derived the excess risk convergence rate of sparse ReLU neural networks estimating piecewise smooth functions, showing that deep learning can outperform the classical linear estimators, including kernel ridge regressors, Fourier estimators, splines, and Gaussian processes. (Here, we refer to the estimators as linear when they are linearly dependent on the output $\mathbf{y}$.) They pointed out that the discrepancy between deep networks and linear estimators appears when the target function is nonsmooth. Suzuki (2018) showed the great adaptiveness of sparse ReLU networks (Equation 2) for the functions in Besov space, a general function space including Hölder space. Specifically, it also allows functions with spatially inhomogeneous smoothness with spikes and jumps. Suzuki (2018) mentioned that deep networks possess strong adaptiveness in capturing the spatial inhomogeneity of functions, whereas linear estimators are only able to capture the global properties of target functions and cannot capture the variability of local shapes. Later, Hayakawa & Suzuki (2020) proved the linear estimators cannot distinguish the function class and its convex hull. This results in the suboptimality of linear estimators over a simple but nonconvex function class, on which sparsely connected deep ReLU nets can attain nearly the minimax optimal rate. There also have been efforts (Farrell et al. 2021, Kohler & Langer 2021) to study the statistical guarantees of densely fully connected networks without sparsity constraints. However, the rates of excess risk they obtained are suboptimal.

2.2.2. Avoiding the curse of dimensionality. According to the classical result from Donoho & Johnstone (1998), for estimating functions in the Hölder class with smoothness $r \geq 0$, the unimprovable minimax convergence rate of excess risk is

$$\inf_{\widehat{f}_n} \sup_{f_\rho \in \{\text{Hölder}\}} \mathbb{E}_{\mathbf{X}\sim\rho_X}\left[(\widehat{f}_n(\mathbf{X}) - f_\rho(\mathbf{X}))^2\right] = \mathcal{O}\left(n^{-2r/2r+d}\right). \quad 7.$$

This rate can be problematic when the data dimension is much larger than the smoothness of function space. In this case, the convergence rate in Equation 7 becomes quite slow in n. Nonetheless, high-dimensional data are often observed in real-world applications. For instance, in the 2012 ImageNet challenge, data were RGB images with a resolution of 224×224, which means $d = 3 \times 224 \times 224$. Then, the rate in Equation 7 cannot explain the empirical success of deep learning. Motivated by this, many researchers have put in considerable effort to avoid the d-dependence in the denominator of the rate in Equation 7.

Several routes exist to avoid the curse. One typical approach is to consider the various types of function spaces $\mathcal{G}$ with various smoothness: mixed-Besov space (Suzuki 2018) and Korobov space (Montanelli & Du 2019). Another alternative is to impose a structural assumption on the target function f_ρ. Such structures include additive ridge functions (Fang & Cheng 2023), composite functions with hierarchical structures (Schmidt-Hieber 2020, Han et al. 2022), generalized single index models (Bauer & Kohler 2019), and multivariate adaptive regression splines (Kohler et al. 2022). Another line of work focuses on the geometric structure of the feature space $\mathcal{X}$. These

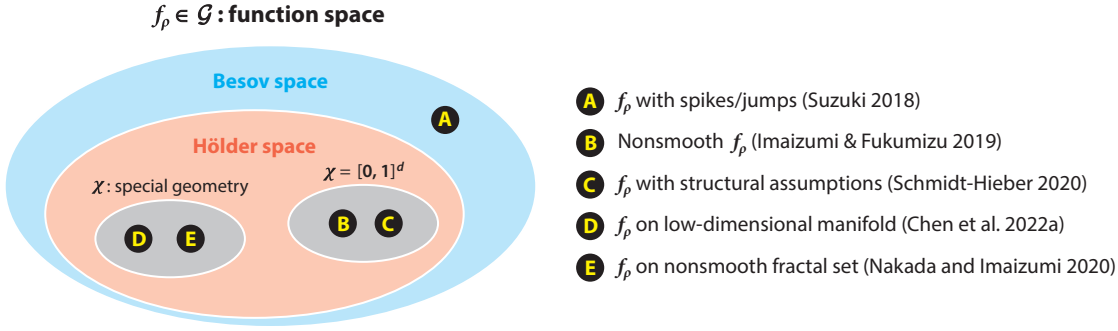

Figure 1

Compared with classical linear estimators (wavelets, kernel ridge regressors, etc.), sparsely connected neural networks are more adaptive in estimating functions f_ρ with special structures. The figure illustrates the different settings of function classes $\mathcal{G}$ where neural networks exhibit superior adaptability over classical estimators.

works take the advantage of high-dimensional data having practically low intrinsic dimensionality (Roweis & Saul 2000, Tenenbaum et al. 2000). Under this setting, Nakada & Imaizumi (2020) showed deep neural networks can achieve a fast rate over a broad class of measures on $\mathcal{X}$, such as data on highly nonsmooth fractal sets. This should be contrasted with the fact that linear estimators, which are known to be adaptive to intrinsic dimensions, can achieve fast convergence rates only when the data lie on smooth manifolds. Chen et al. (2022a) showed the adaptiveness of deep ReLU networks to the data with low-dimensional geometric structures. They were interested in estimating target function f_ρ in Hölder spaces defined over a low-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$. **Figure 1** summarizes the cases on function classes $\mathcal{G}$ where neural nets exhibit superior adaptability over classical estimators.

Recently, Suh et al. (2022) studied deep ReLU networks estimating Hölder functions on a unit sphere and showed that these networks can avoid the curse of dimensionality as function smoothness increases with the data dimension d, $r = \mathcal{O}(d)$. This behavior was not observed in the aforementioned literature, where $\mathcal{X}$ is set as a cube, $\mathcal{X} := [0, 1]^d$. When $\mathcal{X}$ is a cube, several studies (Jiao et al. 2021, Lu et al. 2021, Shen et al. 2021) have attempted to track and reduce the d-dependence in the constant factor hidden in the big-O notation in Equation 7. Interested readers are directed to the detailed comparisons of the results from these works presented by Suh et al. (2022, appendix C).

Note that the aforementioned works are based on the constructions of sparse networks. From a technical perspective, the sparsity assumption is natural in the sense of Equation 6. Nonetheless, as mentioned by Ghorbani et al. (2020), it is still an open question whether the sparsity ($\mathcal{N}$) is a sufficient complexity measure of $\mathcal{F}$ for generalizability, as densely connected networks are observed more commonly in practice without the regularized penalties. This often leads to overparametrized networks with huge complexity on $\mathcal{F}$, which does not guarantee good generalizability in the sense of Equation 6. Given this observation, one popular heuristic argument for explaining the good generalizability of dense neural nets is the implicit regularization of gradient-based algorithms; that is, the model complexity is not captured by an explicit penalty but by the dynamics of the algorithms implicitly. For some special cases (Gunasekar et al. 2018, Ji & Telgarsky 2019), it has been shown that the gradient-based methods provably find the solutions of low complexity in the huge parameter space (e.g., a low-rank matrix in matrix estimation problems). A similar phenomenon has been empirically observed in function estimation problems via neural networks (Cao & Gu 2019, W. Hu et al. 2020), sparking further research that is

discussed in the next section. For more in-depth discussions on these issues beyond what is covered in this review, readers should consult He & Tao (2021, section 3), Neyshabur (2017), and references therein. Reviews on approximation-based statistical guarantees of neural networks for classification problems are provided in **Supplemental Appendix B** due to space limitations.

Supplemental Material >

3. TRAINING DYNAMICS–BASED STATISTICAL GUARANTEES

The literature introduced in Section 2 relies on the assumption that the global minimizer of the empirical risk, $\widehat{f}_n$ in Equation 5, is obtainable. However, due to the nonconvex nature of the loss function, neural networks estimated using commonly employed gradient-based methods lack guarantees of finding $\widehat{f}_n$, which leads to the following natural question: Does the neural network estimated by gradient-based methods generalize well?

The articles discussed shortly try to answer to the above question. Due to the complex nature of the problem, most articles we review consider the following shallow neural network (i.e., a network with one hidden layer) $f_{\mathbf{W}}(\mathbf{x})$ with a number of hidden neurons M:

$$f_{\mathbf{W}}(\mathbf{x}) = \frac{\alpha}{M} \sum_{r=1}^{M} a_r \sigma(w_r^\top \mathbf{x}), \tag{8.}$$

where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ is an input vector, $\{w_r \in \mathbb{R}^d\}_{r=1}^M$ are the weights in the first hidden layer, and $\{a_r \in \mathbb{R}\}_{r=1}^M$ are the weights in the output layer. Let us denote the pair $\mathbf{W} := \{(a_r, w_r)\}_{r=1}^M$. The network dynamic is scaled with the factor $\frac{\alpha}{M}$. If the network width (M) is small, the scaling factor has negligible effects on the network dynamics. But for the wide enough networks (i.e., in the overparametrized setting), the scaling difference yields completely different behaviors in the dynamics. Given a large enough M, we focus on two specific regimes: the neural tangent kernel (NTK) regime (Jacot et al. 2018, Du et al. 2019) with $\alpha = \sqrt{M}$ (Section 3.1) and the MF regime (Mei et al. 2018b, 2019) with $\alpha = 1$ (Section 3.2). Additionally, we review several works that try to address the drawbacks of the NTK framework, as well as some that provide unifying perspectives on these two regimes (Section 3.3).

We focus on reviewing the articles on the ℓ_2-loss function: $\mathcal{L}_{\mathbf{S}}(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^n \left(y_i - f_{\mathbf{W}}(\mathbf{x}_i)\right)^2$. Note that in contrast to Equation 5, structural requirements such as $\mathcal{F}(L, \mathbf{p}, \mathcal{N})$ are removed. The model parameter pairs $\mathbf{W}$ are updated through the gradient-based methods. Let $\mathbf{W}_{(0)}$ be the initialized weight pairs. Then, we have the following GD update rule with step size $\eta > 0$ and $k \geq 1$:

$$\textbf{GD}: \quad \mathbf{W}_{(k)} = \mathbf{W}_{(k-1)} - \eta \nabla_{\mathbf{W}} \mathcal{L}_{\mathbf{S}}(\mathbf{W}) \mid_{\mathbf{W}=\mathbf{W}_{(k-1)}}. \tag{9.}$$

Another celebrated gradient-based method is stochastic GD (SGD). This algorithm takes a randomly sampled subset ($\mathcal{B}$) of the data $\mathcal{D}$ and computes the gradient with the selected samples, and this significantly reduces the computational burdens in GD. Another frequently adopted algorithm in practice is noisy GD, which adds centered Gaussian noise to the gradient of the loss function in Equation 9. Adding noise to the gradient helps with the training (Neelakantan et al. 2015) and generalization (Smith et al. 2020) of neural networks.

3.1. Neural Tangent Kernel Perspective

Over the past few years, the NTK (Chizat & Bach 2018, Jacot et al. 2018, Lee et al. 2018, Arora et al. 2019a) has been one of the most seminal discoveries in the theory of neural networks. The underpinning of the NTK-type theory comes from the observation that in a wide enough neural net, model parameters updated by GD stay close to their initializations during the training, so that the dynamics of the networks can be approximated by the first-order Taylor expansion with

respect to its parameters at initialization. That is, if we denote the output of a neural network as $f_{\mathbf{W}_{(k)}}(\mathbf{x}) \in \mathbb{R}$ with input $\mathbf{x} \in \mathcal{X}$ and model parameter $\mathbf{W}_{(k)}$ updated at the kth iteration of GD, then the dynamics of $f_{\mathbf{W}_{(k)}}(\mathbf{x})$ over $k \geq 1$ can be represented as follows:

$$f_{\mathbf{W}_{(k)}}(\mathbf{x}) = f_{\mathbf{W}_{(0)}}(\mathbf{x}) + \langle \nabla f_{\mathbf{W}_{(0)}}(\mathbf{x}), \mathbf{W}_{(\mathbf{k})} - \mathbf{W}_{(0)} \rangle + o(\|\mathbf{W}_{(\mathbf{k})} - \mathbf{W}_{(0)}\|_{\mathrm{F}}^2), \quad 10.$$

where $o(\|\mathbf{W}_{(\mathbf{k})} - \mathbf{W}_{(0)}\|_{\mathrm{F}}^2)$ is the small random quantity that tends to 0 as the network width gets close to infinity, measuring the distance between an updated model parameter and its initialization in Frobenius norm. Specifically, it can be shown that $\|\mathbf{W}_{(\mathbf{k})} - \mathbf{W}_{(0)}\|_{\mathrm{F}}^2 \leq \mathcal{O}(\frac{1}{\sqrt{M}})$ with sufficiently large M (see, e.g., Du et al. 2018, remark 3.1). In this setting, the right-hand side of Equation 10 is linear in the network parameter $\mathbf{W}_{(k)}$. As a consequence, training on ℓ_2-loss with GD leads to a kernel regression solution with respect to the (random) kernel induced by the feature mapping $\phi(\mathbf{x}) := \nabla_{\mathbf{W}_0} f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. The inner product of two feature mappings evaluated at two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is denoted as $\mathbf{K}^{(M)}(\mathbf{x}_i, \mathbf{x}_j) := \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ for all $1 \leq i, j \leq n$.

Note that $\mathbf{K}^{(M)}(\cdot, \cdot)$ is a random matrix with respect to the initializations $\mathbf{W}_{(0)}$. It has been shown to converge to its deterministic limit ($M \to \infty$) in probability pointwisely (Jacot et al. 2018, Lee et al. 2018, Arora et al. 2019a) and uniformly (Lai et al. 2023) over $\mathcal{X} \times \mathcal{X}$. The limit matrix is named NTK, denoted as $\{\mathbf{K}^{\infty}(\mathbf{x}_i, \mathbf{x}_j)\}_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$. Hereafter, we write the eigen decomposition of $\mathbf{K}^{\infty} = \sum_{j=1}^{n} \lambda_j \mathbf{v}_j \mathbf{v}_j^{\top}$, where $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ with corresponding eigenvectors $\mathbf{v}_j \in \mathbb{R}^n$.

3.1.1. Optimization of neural nets in the neural tangent kernel regime. Many articles tackle the optimization properties of neural networks in the NTK regime. Under the above setting, Du et al. (2018) proved the linear convergence of training loss of shallow ReLU networks. Specifically, the authors randomly initialized $a_r \sim \mathrm{Unif}\{-1, +1\}$ and $\mathbf{w}_r \sim \mathcal{N}(0, \mathcal{I})$, and trained the $\mathbf{w}_r$ via GD with a constant positive step size $\eta = \mathcal{O}(1)$. Here, the linear convergence rate means that the training loss at the kth GD decays at a geometric rate with respect to the initial training loss, which is explicitly stated by Du et al. (2018, theorem 4.1) as

$$\|f_{\mathbf{W}_{(k)}}(\mathbf{x}) - \mathbf{y}\|_2^2 \leq \left(1 - \frac{\eta \lambda_n}{2}\right)^k \|f_{\mathbf{W}_{(0)}}(\mathbf{x}) - \mathbf{y}\|_2^2. \quad 11.$$

Their result requires the network width M to be on the order of $\Omega(\frac{n^6}{\lambda_n})$, and the decay rate is dependent on the minimum eigenvalue of the NTK, λ_n. Here, for the geometric decay rate, λ_n needs to be strictly greater than 0 induced from the data nonparallel assumption (i.e., no two inputs are parallel).

Afterwards, there have been several attempts to reduce the overparametrization size. One work we are aware of is Song & Yang (2020) where they used matrix Chernoff bound to reduce the width size up to $M = \Omega(\frac{n^2}{\lambda_n^4})$ with a slightly stronger assumption than the data nonparallel assumption. Several subsequent works by Allen-Zhu et al. (2019b), Du et al. (2019), Zou et al. (2018), Wu et al. (2019), Oymak & Soltanolkotabi (2020), and Suh et al. (2021) extended the results showing the linear convergence of training loss of deep ReLU networks with L hidden layers. For a succinct comparison of the overparametrized conditions on M in aforementioned articles, we direct readers to Zou & Gu (2019, table 1).

3.1.2. Spectral bias of shallow ReLU networks. Motivated by the result in Equation 11, researchers further studied the spectral bias of deep neural networks, investigating why the neural dynamics learn the lower-frequency components of the functions faster than they learn the higher-frequency counterparts. The specific results are stated in terms of eigenvalues $\mu_1 \geq \mu_2 \geq \ldots$ and corresponding orthonormal eigenfunctions $\phi_1(\cdot), \phi_2(\cdot), \ldots$ of integral operator $\mathcal{L}_{\mathbf{K}^{\infty}}$ induced by $\mathbf{K}^{\infty}$:

$$\mathcal{L}_{\mathbf{K}^{\infty}}(f)(\mathbf{x}) := \int_{\mathcal{X}} \mathbf{K}^{\infty}(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \rho(\mathbf{dy}), \quad \forall f \in \mathcal{L}^2(\mathcal{X}),$$

where $\mathcal{L}^2(\mathcal{X})$ is an L_2-space on $\mathcal{X}$. Specifically, Cao et al. (2020) and Bietti & Mairal (2019) provided the spectral decay rates of $(\mu_k)_k$ for shallow ReLU networks when $\mathbf{x}$ is from a unit-sphere equipped with uniform measure as follows.[1]

Proposition 1 (Cao et al. 2020, theorem 4.3; Bietti & Mairal 2019, proposition 5). For the NTK corresponding to a two-layer feed-forward ReLU network, the eigenvalues $(\mu_k)_k$ satisfy the following:

$$\begin{cases} \mu_k = \Omega(1), & \text{when } k = 0, 1, \\ \mu_k = 0, & \text{when } k(\geq 3) \text{ is odd}, \\ \mu_k = \Omega(\max(k^{-d-1}, d^{-k-1})), & \text{when } k(\geq 2) \text{ is even}. \end{cases}$$

The decay rate is exponentially fast in input dimension d for $k \gg d$. An interesting benefit of having a specific decay rate is that we can measure the size of RKHSs induced from the kernel $\mathbf{K}^\infty$. (We denote this RKHS as $\mathcal{H}^\infty$ for later use.) The slower the decay rate is, the larger the RKHS becomes, allowing higher-frequency information of function to be included.

With the specified eigendecay rates on μ_k, Cao et al. (2020, theorem 4.2) proved the spectral bias of neural network training in the NTK regime. Specifically, as long as the network is wide enough and the sample size is large enough, GD first learns the target function along the eigendirections of NTK with larger eigenvalues, and learns the other components corresponding to smaller eigenvalues later. Similarly, Hu et al. (2019) showed that GD learns the linear component of target functions in the early training stage. But, crucially, they do not require the network to have a disproportionately large width, and the network is allowed to escape the kernel regime later in training.

3.1.3. Generalization of neural nets in the neural tangent kernel regime. Here, we review some important works that study the generalizability of Equation 8. To the best of our knowledge, Arora et al. (2019b) provided the first step in understanding the role of NTK in the generalizability of neural nets. Specifically, they showed that for $M = \Omega(\frac{n^2 \log(n)}{\lambda_n})$ and $k \geq \tilde{\Omega}(\frac{1}{\eta \lambda_n})$, the ℓ_2-population loss of $f_{\mathbf{W}_{(k)}}(\mathbf{x})$ is bounded by

$$\mathbb{E}\big[(f_{\mathbf{W}_{(k)}}(\mathbf{x}) - \mathbf{y})^2\big] \leq \mathcal{O}\left(\sqrt{\frac{\mathbf{y}^\top (\mathbf{K}^\infty)^{-1} \mathbf{y}}{n}}\right). \tag{12.}$$

Observe that the numerator in the bound can be written as $\mathbf{y}^\top (\mathbf{K}^\infty)^{-1} \mathbf{y} := \sum_{i=1}^n \frac{1}{\lambda_i} (\mathbf{v}_i^\top \mathbf{y})^2$. This implies the projections $\mathbf{v}_i^\top \mathbf{y}$ that correspond to small eigenvalues λ_i should be small for good generalizations on unseen data. This theoretical result is consistent with the empirical finding of Zhang et al. (2018), who performed empirical experiments on MNIST and CIFAR-10 datasets, showing that the projections $\{(\mathbf{v}_i^\top \mathbf{y})\}_{i=1}^n$ sharply drop for true labels $\mathbf{y}$, leading to the fast convergence rate. In contrast, when the projections are close to being uniform for random labels $\mathbf{y}$, it leads to slow convergence (see Zhang et al. 2018, figure 1).

However, the bound of Equation 12 is obtained in the noiseless setting and becomes vacuous under the presence of noise. In this regard, under the noisy setting (Equation 4), Nitanda & Suzuki (2020) showed that

$$\mathbb{E}\big[\| f_{W_{(k)}}(\mathbf{x}) - f_\rho(\mathbf{x}) \|_{L_2}^2\big] \leq \mathcal{O}\big(k^{-\frac{2r\beta}{2r\beta+1}}\big), \tag{13.}$$

[1] Note that eigenvalues of $\mathbf{K}^\infty$ ($\{\lambda_i\}_{i=1}^n$) and eigenvalues of $\mathcal{L}_k$ [$(\mu_k)_k$] are different.

where the target function f_ρ belongs to the subset of $\mathcal{H}^\infty$, and $f_{W_{(k)}}$ is a shallow neural network with a smooth activation function that approximates ReLU. Here, the network is estimated via one-pass SGD (take one sample for gradient update and the samples are visited only once during training), minimizing ℓ_2-regularized expected loss. This setting leads to $k = n$. The rate of Equation 13 is minimax optimal, which is faster than $\mathcal{O}(\frac{1}{\sqrt{n}})$ from Arora et al. (2019b). It is characterized by two control parameters, β and r, where $\beta > 1$ controls the size of $\mathcal{H}^\infty$ and $r \in [1(/)2, 1]$ controls the size of subset of $\mathcal{H}^\infty$ where f_ρ belongs. The bound of Equation 13 has an interesting bias–variance trade-off between these two quantities β and r. For large β, the whole space $\mathcal{H}^\infty$ becomes small, and the subspace of $\mathcal{H}^\infty$ needs to be as large as possible for the faster convergence rate, and vice versa.

However, as noted by Hu et al. (2021), the rate in Equation 13 requires the network width M to be exponential in n. The work reduced the size of overparametrization to $\tilde{\Omega}(n^6)$ when the network parameters are estimated by GD. The article proved that the overparametrized shallow ReLU networks require ℓ_2-regularization for GD to achieve the minimax convergence rate $\mathcal{O}(n^{-\frac{d}{2d-1}})$. Later, Suh et al. (2021) extended the result to deep ReLU networks in the NTK regime, showing that ℓ_2-regularization is also required for achieving the minimax rate for deep networks.

3.2. Mean-Field Perspective

A MF viewpoint is another interesting paradigm to help us understand the optimization landscape of neural network models. Recall that neural network dynamics in the MF regime corresponds to $\alpha = 1$ in Equation 8.

The term mean-field comes from an analogy with mean-field models in mathematical physics, which analyze the stochastic behavior of many identical particles (Ryzhik 2023). Let us denote $\theta_r := (a_r, w_r) \in \mathbb{R}^{d+1}$ and $\sigma_\star(\mathbf{x}, \theta_r) := a_r \sigma(w_r^\top \mathbf{x})$ in Equation 8. Weight pairs $\{\theta_r\}_{r=1}^M$ are considered as a collection of gas particles in $\mathbf{D}$-dimensional spaces with $\mathbf{D} := d + 1$. We consider that there are infinitely many gas particles, allowing $M \to \infty$, which yields the following integral representation of neural dynamics:

$$\frac{1}{M}\sum_{r=1}^{M} \sigma_\star(\mathbf{x}; \theta_r) \xrightarrow{M\to\infty} f(\mathbf{x}; \rho) := \int \sigma_\star(\mathbf{x}; \theta)\rho(\mathrm{d}\theta), \qquad 14.$$

where $\theta_r \sim \rho$ for $r = 1, \ldots, M$. The integral representation of Equation 14 is convenient for the mathematical analysis as it is linear with respect to the measure ρ (see, e.g., Bengio et al. 2005).

Under this setting, the seminal work of Mei et al. (2018b) studied the evolution of particles $\theta^{(k)} \in \mathbb{R}^{\mathcal{D}}$ updated by k steps of one-pass SGD (take one sample for the gradient update, and the samples are visited only once during training) under ℓ_2-loss. Interestingly, they proved that the trajectories of the empirical distribution of $\theta^{(k)}$, denoted as $\widehat{\rho}_k := \frac{1}{M}\sum_{r=1}^{M} \delta_{\theta_r^{(k)}}$, weakly converge to the deterministic limit $\rho_t \in \mathcal{P}(\mathbb{R}^{\mathcal{D}})$ as $k \to \infty$ and $M \to \infty$. The measure ρ_t is the solution of the following nonlinear partial differential equation (PDE):

$$\partial_t \rho_t = \nabla_\theta \cdot \big(\rho_t \nabla_\theta \boldsymbol{\Psi}(\theta; \rho_t)\big), \qquad \boldsymbol{\Psi}(\theta; \rho_t) := \mathcal{V}(\theta) + \int \mathcal{U}(\theta, \bar{\theta})\rho_t(\mathrm{d}\bar{\theta}), \qquad 15.$$

$$\mathcal{V}(\theta) := -\mathbb{E}\{\mathbf{y}\sigma_\star(\mathbf{x}; \theta)\}, \qquad \mathcal{U}(\theta_1, \theta_2) := \mathbb{E}\{\sigma_\star(\mathbf{x}; \theta_1)\sigma_\star(\mathbf{x}; \theta_2)\}.$$

The above PDE describes the evolution of each particle (θ_r) in the force field created by the densities of all the other particles. (We provide more descriptions of the above PDE in **Supplemental Appendix C**.) Denote $\mathcal{R}(\rho_t) := \mathbb{E}[(y - f(\mathbf{x}; \rho_t))^2]$ and let $\mathcal{R}_M$ be the empirical

Supplemental Material >

version of $\mathcal{R}$; then, under some regularity assumptions on the network, we have

$$\sup_{0\le t\le T}\left|\mathcal{R}(\rho_t)-\mathcal{R}_M(\widehat{\rho}_{\lfloor t/2\eta\rfloor})\right|\le e^{C(T+1)}\cdot\sqrt{\frac{1}{M}\vee 2\eta}\cdot\sqrt{D+\log\frac{M}{2\eta}}. \tag{16.}$$

The conditions for the bound to vanish to 0 are (*a*) $M\gg D$, (*b*) $\eta\ll\frac{1}{D}$, and (*c*) the PDE converges in $T=\mathcal{O}(1)$ iterations. It is interesting to note that the generic ordinary differential equation approximation requires the step size η to be less than the order of the total number of parameters in the model ($\eta\ll\frac{1}{MD}$), whereas in this setting the step size $\eta\ll\frac{1}{D}$ should be enough. Also, recall that the number of sample size n is equivalent to the iteration steps $k:=\lfloor\frac{T}{\varepsilon}\rfloor$ of one-pass SGD with $T=\mathcal{O}(1)$. Then, this means $n=\mathcal{O}(D)\ll\mathcal{O}(MD)$ should be enough for a good approximation. Another notable fact is that, in contrast to the NTK regime, the evolution of weights θ_r is nonlinear, and in particular, the weights move away from their initialization during training. Indeed under mild assumptions, we can show that for a small enough step size η, $\lim_{M\to\infty}\|\theta^{(k)}-\theta^{(0)}\|_2^2/M=\Omega(\eta^2)$ in the MF regime, while $\sup_{t\ge 0}\|\theta^{(t)}-\theta^{(0)}\|_2^2/M=\mathcal{O}(n/(Md))$ in the NTK regime (see Bartlett et al. 2021).

Despite the nice characterizations of SGD dynamics, the bound in Equation 16 still has room for improvement; the number of neurons M is dependent on the ambient data dimension d, and the bound is only applicable to the SGD with short convergence iterations $T=\mathcal{O}(1)$. A follow-up work (Mei et al. 2019) has attempted to tackle these challenges. Particularly, they proved that there exists a constant K that only depends on intrinsic features of the activation and data distribution, such that with high probability, the following holds:

$$\sup_{0\le t\le T}\left|\mathcal{R}(\rho_t)-\mathcal{R}_M(\widehat{\rho}_{\lfloor t/\varepsilon\rfloor})\right|\le Ke^{K(T\eta)^3}\left\{\sqrt{\frac{\log(M)}{M}}+\sqrt{d+\log(M)}\sqrt{\eta}\right\}. \tag{17.}$$

A remarkable feature of this bound is that as long as $T\eta=\mathcal{O}(1)$ and $K=\mathcal{O}(1)$, the number of neurons only needs to be chosen where $M\gg 1$ for the MF approximation to be accurate. The condition $T\eta=\mathcal{O}(1)$ mitigates the exponential dependence on T, and the bound does not need to scale with the ambient dimension d. Later, researchers from the same group generalized the result into multi-layer settings (Nguyen & Pham 2023).

3.3. Beyond the Neural Tangent Kernel and Mean-Field Regimes

Despite nice theoretical descriptions on the training dynamics of GD in loss functions, Arora et al. (2019a), Lee et al. (2018), and Chizat & Bach (2018) empirically found significant performance gaps between NTK and actual training in many downstream tasks. For instance, Arora et al. (2019a) derived the CNN-based convolutional NTK (CNTK) and empirically found 5% $\sim$ 6% performance gaps between the CNN- and CNTK-based kernel regressors in image classification tasks, with CNN performing better. This indicates the potential benefits of finite width in neural networks.

3.3.1. Beyond the neural tangent kernel regime. These gaps have been theoretically studied in several articles, including those of Wei et al. (2019), Allen-Zhu & Li (2019), Ghorbani et al. (2021a), Yehudai & Shamir (2019), and Chizat & Bach (2018). They showed that NTK has worse generalization guarantees than finite-width neural networks in some settings. For example, Mei et al. (2018a) demonstrated that training a neural network with one hidden neuron means that it can efficiently learn a single neuron target function with $\mathcal{O}(d\log d)$ samples, whereas the corresponding RKHS has a test error that is bounded away from zero for any sample size polynomial in d (Yehudai & Shamir 2019, Ghorbani et al. 2021b). However, kernel methods often perform comparably to neural networks in some image classification tasks (Li et al. 2019, Novak et al.

2020). Ghorbani et al. (2021a) provided a unified framework under spiked covariate models to explain this divergence, showing that while RKHSs and neural networks perform similarly in certain stylized tasks, RKHS performance deteriorates under isotropic covariate distributions (e.g., noisy high-frequency image components), whereas neural networks are less affected by such noise. Wei et al. (2019) gave an interesting example in which NTK or any kernel methods are statistically limited, whereas regularized neural networks have better sample complexity. They proved that there is a $\Omega(d)$ sample-complexity gap between the regularized neural net and kernel prediction function for estimating $f_\rho(\mathbf{x}) = \mathbf{x}_1\mathbf{x}_2$ with $\mathbf{x}_i \sim \{\pm 1\}$ for $\mathbf{x} \in \mathbb{R}^d$.

The aforementioned works explained the superiority of neural networks over the networks in the NTK regime under some highly stylized settings. There also has been another line of works (Li & Liang 2018, Allen-Zhu et al. 2019a, Bai & Lee 2020) to explain how the networks estimated through gradient-based methods generalize well but, critically, do not rely on the linearization of network dynamics. Under the distribution-free setting (i.e., no distributional assumptions on covariates), Allen-Zhu et al. (2019a) provided optimization and generalization guarantees for three-layer ReLU networks, learning the function classes of three-layer networks with smooth activation functions. They showed that three-layer ReLU networks can learn a larger function class than two-layer ReLU networks do. Unlike NTK techniques, their approach allows nonconvex interactions across hidden layers, enabling parameters trained by SGD to move far from their initializations. Motivated from Allen-Zhu et al. (2019a), Bai & Lee (2020) studied the optimization and generalization of shallow networks with smooth activation functions $\sigma(\cdot)$ via relating the network dynamics $f_W(\mathbf{x})$ in Equation 8 with second-order (or quadratic) approximations. They explicitly showed that the sample complexity of the quadratic model is smaller than that of the linear NTK model in learning some stylized target functions by the factor of $\mathcal{O}(d)$. Similarly, relying on tensor decomposition techniques instead of quadratic approximation, Li & Liang (2018) showed the separations of shallow ReLU networks and NTK regressors.

3.3.2. Unifying views of neural tangent kernel and mean-field regimes. There have been several attempts to give a unifying view of the NTK and MF regimes, including that of Chen et al. (2020). The article is motivated by complementing the cons of both regimes, whose pros and cons are summarized in **Table 1**. Chen et al. (2020) showed the two-layer neural networks learned through noisy GD in MF scaling can potentially learn a larger class of functions than networks in NTK scaling can do. One seminal work, that of Yang & Hu (2022), identified a set of scales for initialized weights and SGD step sizes where feature learning occurs in deep neural networks in the infinite width limit. They offer a unified framework that encompasses parametrizations in both the NTK and MF regimes. Feature learning is the core property of neural networks: the ability to learn useful features out of raw data (Girshick et al. 2014, Devlin et al. 2018) that adapt to the learning problem. For instance, BERT (bidirectional encoder representations from transformers) (Devlin et al. 2018) leveraged this property of neural networks for sentence sentiment analysis.

Table 1 Neural tangent kernel versus mean-field regimes

	Neural tangent kernel regime	Mean-field regime
Pros	1. Same scaling as in practice	1. Does not require $\theta^{(k)}$ to be close to $\theta^{(0)}$
	2. Finite time convergence rate	2. Potentially learns a larger class of functions
	3. Generalization bounds	
Cons	1. Requires $\theta^{(k)}$ to be close to $\theta^{(0)}$	1. Not the same scaling as in practice
		2. No finite-time convergence rate
		3. No generalization bounds

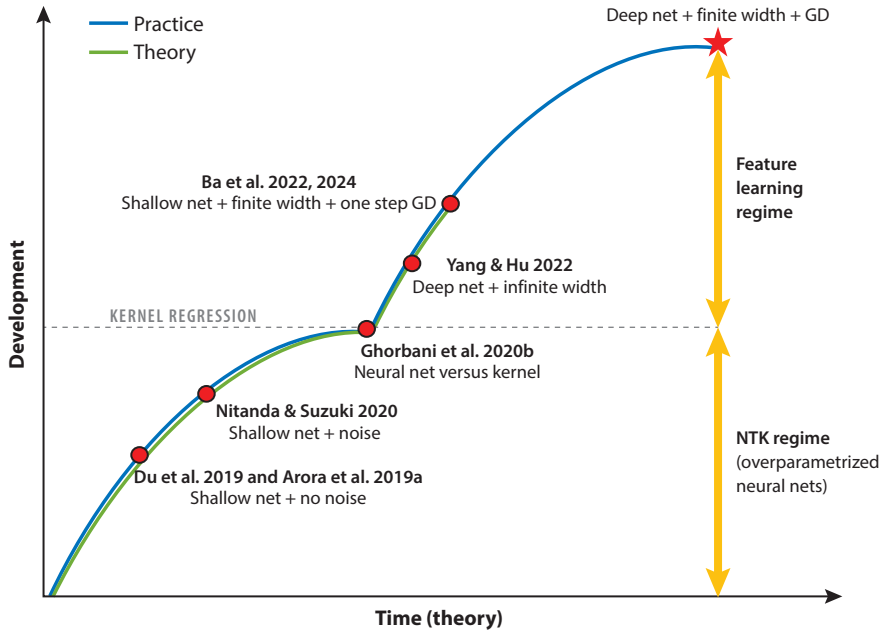

Figure 2

Development of the literature (*y*-axis) on algorithm-based neural network analysis over time (*x*-axis). In our view, the ultimate goal (represented as a star) of this line of research is to theoretically demystify feature learning of neural nets with deep layers and finite widths, closing the gap between theory and practice. Note that the kernel regressor in the NTK regime does not exhibit feature learning functionality. Abbreviations: GD, gradient descent; NTK, neural tangent kernel.

Specifically, in the regime where the network width and data size are comparable, Ba et al. (2022) showed that nontrivial feature learning occurs at the early phase (one gradient step in GD) of shallow neural network training with a large enough step size. Similarly to Ghorbani et al. (2021a), Ba et al. (2024) examined the advantages of shallow neural networks with finite width over kernel methods under the spiked covariance model. Both studies demonstrated that neural networks and kernel methods benefit from stronger low-dimensional structures (i.e., larger spikes). However, Ba et al. (2024) focused on gradient-based optimization guarantees, while Ghorbani et al. (2021a) provided only approximation-based analysis.

Supplemental Material >

Interested readers can find more detailed descriptions on the works by Wei et al. (2019), Ghorbani et al. (2021a), Bai & Lee (2020), Chen et al. (2020), and Yang & Hu (2022) in **Supplemental Appendixes D** and **E**. In **Figure 2**, we summarize our own views on the developments in the literature along this line of research.

4. STATISTICAL GUARANTEES OF GENERATIVE MODELS

In this section, we sequentially explore the statistical literature on three topics: GANs, diffusion models, and ICL in LLMs.

4.1. Generative Adversarial Networks

Over the past decade, GANs (Goodfellow et al. 2014) have stood out as a significant unsupervised learning approach and are known for their ability to learn the data distribution and efficiently sample the data from it. The main goal of GANs is to learn the target distribution $\mathbf{X} \sim \nu$ through

adversarial training between a discriminator and a generator. Here, the generator takes the input $\mathcal{Z}$ from prior distributions such as Gaussian or uniform ($\mathcal{Z} \sim \pi$) and the input is push-forwarded by the transformation map $g : \mathcal{Z} \to g(\mathcal{Z})$. In the seminal article of Goodfellow et al. (2014), the distribution of random variable $g(\mathcal{Z})$ is referred to as an implicit distribution μ. The primary objective of the generator is to produce synthetic data from μ that closely resembles samples from the target distribution.

The adversarial training between the discriminator and generator is enabled through optimizing the following minimax problem with respect to functions in the generator class $\mathcal{G}$ and discriminator class $\mathcal{F}$:

$$(g^\star, f^\star) \in \arg\min_{g \in \mathcal{G}} \arg\max_{f \in \mathcal{F}} \left\{ \mathbb{E}_{\mathcal{Z} \sim \pi} f(g(\mathcal{Z})) - \mathbb{E}_{X \sim \nu} f(X) \right\}. \quad 18.$$

In practice, the above expectations can be approximated with m training data from ν and n drawn samples from μ. The inner maximization problem in Equation 18 is an integral probability metric (IPM) (Müller 1997), which quantifies the discrepancy between two distributions μ and ν with respect to a symmetric function class, $f \in \mathcal{F}$, then $-f \in \mathcal{F}$, with

$$d_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_{x \sim \mu} f(x) - \mathbb{E}_{y \sim \nu} f(y) \right\}. \quad 19.$$

When $\mathcal{F}$ is taken to be all 1-Lipschitz functions, $d_{W_1}(\cdot, \cdot)$ is the Wasserstein-1 distance; when $\mathcal{F}$ is the class of all indicator functions, $d_{\mathrm{TV}}(\cdot, \cdot)$ is the total variation (TV) distance; and when $\mathcal{F}$ is taken as a class of neural networks, $d_{\mathrm{NN}}(\cdot, \cdot)$ is the neural net distance (see Arora et al. 2017). Under this setting, a generator from $\mathcal{G}$ attempts to minimize the IPM between μ and ν.

4.1.1. Generalization of generative adversarial networks. A question that naturally arises is, What does it mean for GANs to generalize effectively? Arora et al. (2017) provided a mathematical definition for generalization in GANs in terms of IPM.

Definition 1. Let $\widehat{\mu}_n$ and $\widehat{\nu}_m$ be the empirical distributions of μ and ν. For some generalization gap $\varepsilon > 0$, if it holds with high probability that

$$\left| d_{\mathcal{F}}(\mu, \nu) - d_{\mathcal{F}}(\widehat{\mu}_n, \widehat{\nu}_m) \right| \leq \varepsilon, \quad 20.$$

with n being polynomially dependent in ε, then the divergence $d_{\mathcal{F}}(\cdot, \cdot)$ between distributions generalizes.

This means that if the absolute discrepancy between population divergence and empirical divergence of μ and ν can be arbitrarily controlled with n polynomially generated samples, the GAN generalizes well. The same article proved, under this definition, that GANs cannot generalize with respect to Wasserstein-1 distance and Jensen-Shannon divergence as $n = \tilde{\mathcal{O}}(\varepsilon^{-\mathrm{poly(d)}})$ is required. But they generalize well with respect to neural net distance with $n = \tilde{\mathcal{O}}(p \log(L) \cdot \varepsilon^{-2})$, where p is the total number of parameters in the discriminator neural network and L is a Lipschitz constant of discriminators with respect to parameters.

Nonetheless, as noted by Chen et al. (2022b), Zhang et al. (2018), and Arora et al. (2017), this result has some limitations: (*a*) The sample complexity is involved with unknown Lipschitz constants L of the discriminator. (*b*) A small neural net distance does not necessarily mean that two distributions are close (Arora et al. 2017, section 3.4). (*c*) Sample complexity is not involved with the complexity of generator class $\mathcal{G}$ under the assumption that the generator can approximate well enough the target data distribution. (*d*) No concrete architecture of discriminator networks is given. Some articles attempted to address the first two limitations (Zhang et al. 2018, Jiang et al. 2018, Bai et al. 2019), and their attempts are nicely summarized by Chen et al. (2022b). In this review, we discuss works by Chen et al. (2022b) and Liang (2021) that tackle the issues raised in points *c* and *d* concretely through tools from the approximation theory of neural networks.

4.1.2. Statistical guarantees of generative adversarial networks. Note that the functions in discriminator and generator classes either can be classical nonparametric regressors (random forests, local polynomial, etc.) or can be both parametrized by neural networks. Here, we focus on the latter case, which is more commonly used in practice. Specifically, let us denote $\mathcal{F} := \{f_\omega(\cdot) : \mathbb{R}^d \to \mathbb{R}\}$ as the discriminator class and $\mathcal{G} := \{g_\theta(z) : \mathbb{R}^m \to \mathbb{R}^d\}$ (with $m \le d$) as the generator class, with ω and θ being denoted as network parameters of the respective classes, and we are interested in estimating the parameters by minimizing the following optimization problem:

$$(\widehat{\theta}_{m,n}, \widehat{\omega}_{m,n}) \in \underset{\theta: g_\theta \in \mathcal{G}}{\arg\min} \max_{\omega: f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega(g_\theta(\mathcal{Z})) - \widehat{\mathbb{E}}_n f_\omega(X) \right\}, \tag{21.}$$

where $\widehat{\mathbb{E}}_m(\cdot)$ [resp. $\widehat{\mathbb{E}}_n(\cdot)$] denotes the empirical expectation with m generator samples (resp. n training samples).

4.1.3. Summary of Liang (2021). Given that the optimal parameters of generator $\widehat{\theta}_{m,n}$ in Equation 21 can be obtained, Liang (2021) studied how well the implicit distribution estimator $\mu_{\widehat{\theta}_{m,n}}$ [i.e., the distribution of the random variable $g_{\widehat{\theta}_{m,n}}(Z)$] gets close to the target distribution ν in the TV distance. Under some regularity assumptions on the architectures of g_θ and f_w, Liang (2021, theorem 19) proved the existence of $(g_\theta(z), f_\omega)$ pairs satisfying the bound

$$\mathbb{E} d_{\mathrm{TV}}^2(\nu, \mu_{\widehat{\theta}_{m,n}}) \le \sqrt{d^2 L \log(dL) \left(\frac{\log m}{m} \vee \frac{\log n}{n} \right)}. \tag{22.}$$

In the rate of Equation 22, L and d are the depth and width of the generator networks, respectively. This result allows the very deep network as $L \le \sqrt{(n \wedge m)/\log(n \vee m)}$. It is worth noting that the generator requires the width of the network to be the same as the input dimension d so that the invertibility condition on the generator is satisfied. As for the discriminator, it can be constructed by concatenating two networks that have the same architecture as the one from the generator, and with the additional two layers (i.e., network f_ω has $L + 2$ layers). The g_θ and f_ω used leaky ReLU and dual leaky ReLU as activation functions, respectively, for their invertibility. However, this invertibility condition is often violated in practical uses of GANs.

4.1.4. Summary of Chen et al. (2022b). Chen et al. (2022b) subsequently provided more flexible network architectures for g_θ and f_ω without requiring the invertibility condition on generator and activation functions (i.e., the authors consider a ReLU activation function). The article mainly focuses on three interesting scenarios that impose structural assumptions on the target distribution ν:

1. The target distribution ν is assumed to have a $\alpha(>0)$-Hölder density p_ν with respect to Lebesgue measure in $\mathbb{R}^d$, and the density is lower-bounded away from 0 on a compact convex subset $\mathcal{X} \subset \mathbb{R}^d$.
2. The target distribution ν is supported on the q-dimensional ($q \ll d$) linear subspace of the data domain $\mathcal{X} \subset \mathbb{R}^d$, where the density function is assumed to be in α-Hölder class.
3. The target distribution ν is supported on $\mathcal{X} \subset [0,1]^d$ with q-dimensional ($q \ll d$) nonlinear K-mixture components, where each component's density function is in α-Hölder class.

In scenario 1, the discriminator class $\mathcal{F}$ is assumed to be the β-Hölder class for $\beta > 1$. In scenarios 2 and 3, $\mathcal{F}$ is considered to be a collection of 1-Lipschitz functions. The convergence rate of the IPM, the depth L, and the maximum widths $\mathbf{p}_{\max}$ of the generator and discriminator in each scenario are summarized in **Table 2**.

In the scenario 1, the convergence rate cannot avoid the exponential dependence on d, aligning with the known minimax lower bound (Tang & Yang 2022). The result in scenario 2 indicates

Table 2 Summary of depth (L) and width (p_{max}) of generators (g_θ) and discriminators (f_w) and the convergences of integral probability metrics under three specially designed scenarios in the generative adversarial networks framework

Scenario		L	p_{max}	Convergence rate
Scenario 1	g_θ	$\mathcal{O}\left(\frac{\beta}{2\beta+d}\log n\right)$	$\mathcal{O}\left(dn^{\frac{\beta d}{(\alpha+1)(2\beta+d)}}\right)$	$\tilde{\mathcal{O}}\left(n^{-\frac{\beta}{2\beta+d}}\log^2 n\right)$
	f_ω	$\mathcal{O}\left(\frac{\beta}{2\beta+d}\log n\right)$	$\mathcal{O}\left(n^{\frac{2d}{2\beta+d}}\right)$	
Scenario 2	g_θ	$\mathcal{O}\left(\frac{\alpha}{2\alpha+q}\log n\right)$	$\mathcal{O}\left(qn^{\frac{q\alpha}{(\alpha+1)(2\alpha+d)}}\vee d\right)$	$\tilde{\mathcal{O}}\left(n^{-\frac{1}{2+q}}\log^2 n\right)$
	f_ω	$\mathcal{O}\left(\frac{1}{2+q}\log n\right)$	$\mathcal{O}\left(n^{\frac{q}{2+q}}\vee d\right)$	
Scenario 3	g_θ	$\mathcal{O}\left(\frac{1}{q}\log n\right)$	$\mathcal{O}\left(Kdn^{\frac{1}{\alpha}}\right)$	$\mathcal{O}\left(dn^{-\frac{1}{q}}\right)$
	f_ω	$\mathcal{O}\left(\log n+d\right)$	$\mathcal{O}\left(n^{\frac{d}{q}}\right)$	

GANs can avoid the curse of dimensionality by being adaptive to the unknown low-dimensional linear structures and achieving faster rates than the parametric rate $n^{-\frac{1}{2}}$. However, rather than the real-world data being centered in the low-dimensional linear subspace, mixture data are more commonly observed in practice (e.g., MNIST data or images in CIFAR-10). In the scenario 3, the rate depends linearly on d and exponentially on q, showing that GANs can capture nonlinear data structures. The depth L of networks grows logarithmically with sample size n. In contrast with the work of Liang (2021), the widths of networks are not the same as the input dimension d.

4.2. Score-Based Diffusion Models

Score-based diffusion models consist of two processes (Song et al. 2020). The first step, the forward process, transforms data into noise. Specifically, the score-based diffusion model uses the following stochastic differential equation (SDE) (Särkkä & Solin 2019) for data perturbation:

$$\mathrm{d}\mathbf{x} = f(\mathbf{x},t)\mathrm{d}t + g(t)\mathrm{d}W_t, \qquad 23.$$

where $f(\mathbf{x}, t)$ and $g(t)$ are the drift and diffusion coefficients, respectively, and W_t is a standard Wiener process (a.k.a. Brownian motion) indexed by time $t \in [0, T]$. Here, the f and g functions are user-specified, and Song et al. (2020) suggest three different types of SDEs for data perturbation: variance exploding, variance preserving, and subvariance preserving. Allowing diffusion to continue long enough with T being sufficiently large, it can be shown that the distribution of $\mathbf{x}_t$ converges to some easy-to-sample distributions π, such as normal or uniform distributions. Specifically, when $f := -\mathbf{x}_t$ and $g := \sqrt{2}$, Equation 23 is known as the Ornstein–Uhlenbeck process, and it has been proven that $p_t := \mathrm{Law}(\mathbf{x}_t) \to \pi$, with π being normal and exponentially fast in 2-Wasserstein distance (see, e.g., Bakry et al. 2014). However, despite this convergence result, it is analytically difficult to know the exact form of p_T, and it is often replaced by π in practice when starting the reverse process.

The second step, reverse process, is a generative process that reverses the effect of the forward process. This process learns to transform the noise back into the data by reversing the SDEs in Equation 23. Through the Fokker–Planck equation of marginal density $p_t(\mathbf{x})$ for time $t \in [t_0, T]$, the following reverse SDE (Anderson 1982) can be easily derived:

$$\mathrm{d}\mathbf{x} = \left[f(\mathbf{x},t) - g(t)^2\nabla_{\mathbf{x}}\log p_t(\mathbf{x})\right]\mathrm{d}t + g(t)\mathrm{d}\bar{W}_t. \qquad 24.$$

Here, the gradient of $\log p_t(\mathbf{x})$ with respect to the perturbed data $\mathbf{x}(t)$ is referred to as a score function, $\mathrm{d}t$ in Equation 24 is an infinitesimal negative time step, and $\mathrm{d}\bar{W}_t$ is a Wiener process

running backward in time, with t: $T \to t_0$. In practice, t_0 is usually chosen to be a small number close to 0, but not too close to 0 to prevent the blow up of the score function. There are various ways to solve Equation 24—for instance, a discretization scheme such as Euler–Maruyama, or a theory-driven method such as probability flow (for a more detailed exposition on these methods, see Song et al. 2020). The papers cited in this article focus on the discretization scheme, and readers can refer to S. Chen et al. (2023a) for recent theoretical understanding of the probability flow in the diffusion model.

The score function, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, is approximated by a time-dependent score-based model $\mathbf{S}_\theta(\mathbf{x}(t), t)$ which is parametrized by neural networks in practice. The network parameter θ is estimated by minimizing the following score-matching loss:

$$\theta^\star := \arg\min_\theta \mathbb{E}_{t\sim\mathcal{U}[t_0,T]} \mathbb{E}_{\mathbf{x}(\mathbf{t})|\mathbf{x}(\mathbf{0})} \mathbb{E}_{\mathbf{x}(0)} \left[\lambda(t)^2 \left\| \mathbf{S}_\theta(\mathbf{x}(t), t) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x}(t) \mid \mathbf{x}(0)) \right\|_2^2 \right], \qquad 25.$$

where $\mathcal{U}[t_0, T]$ is a uniform distribution over $[t_0, T]$, and $\lambda(t)(> 0)$ is a positive weighting function that helps the scales of matching losses $\left\| \mathbf{S}_\theta(\mathbf{x}(t), t) - \nabla_{\mathbf{x}} \log p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0)) \right\|_2^2$ to be in the same order over the time $t \in [t_0, T]$. The transition density $p_t(\mathbf{x}(t) \mid \mathbf{x}(0))$ is a tractable Gaussian distribution, and $\mathbf{x}(t)$ can be obtained through ancestral sampling (Ho et al. 2020).

Under this setting, we introduce readers to two lines of attempts to find answers to the following theoretical questions:

1. Can the diffusion model estimate the target distribution ν via the learned score function? If so, under what conditions on ν can we guarantee polynomial convergence in the generalization error bound ε, when measured in terms of TV or Wasserstein distances?
2. Do neural networks well approximate and learn the score functions? If so, how should one choose network architectures, and what is the sample complexity of learning? Furthermore, if the data distribution has a special geometric structure, is the diffusion model adaptive to the structure, like GAN models are?

The main statistical object of interest in these two lines of research is the generalization bound measuring the distance between target distribution ν and estimated distribution $\widehat{\mu}_\theta$ from the samples $\mathbf{x}_{t_0}$ by solving the reverse SDE in Equation 24. Here, the score function is substituted by the estimated time-dependent neural network $\mathbf{S}_\theta(\mathbf{x}(t), t)$. The first line of work mainly focuses on the sampling perspective of the diffusion model, given that we have good estimates of the score function. The second line of work extends the attention to the score function approximation through neural networks. Furthermore, under some highly stylized settings, the second line of work specifies the explicit network structures that give good generalization guarantees.

4.2.1. Attempts to answer question 1. Early theoretical efforts to understand the sampling of score-based diffusion models suffered either from not being quantitative (De Bortoli et al. 2021, Liu et al. 2022) or from the curse of dimensionality (Block et al. 2020, De Bortoli 2023). Specifically, De Bortoli (2023) gave the convergence in the 1-Wasserstein distance for distributions with bounded support $\mathcal{M}$. This case covers the distributions supported on lower-dimensional manifolds, where guarantees in TV or Kullback–Leibler distance are unattainable as there are no guarantees that ν and $\widehat{\mu}_\theta$ have the same support set. For general distributions, their bounds on $W_1(\nu, \widehat{\mu}_\theta)$ have exponential dependence on the diameter of the manifold $\mathcal{M}$ and truncation of the reverse process t_0 as $\mathcal{O}(\exp(\mathrm{diam}(\mathcal{M})^2/t_0))$. For smooth distributions where the Hessian $\nabla^2 \log p_t$ is available, the bound is further improved with a polynomial dependence on t_0 with the growth rate of the Hessian as $t \to 0$ being on the exponent.

To the best of our knowledge, Lee et al. (2022) first gave the polynomial guarantees in TV distance under a L^2-accurate score for a reasonable family of distributions. However, their result

is based on the assumption that the distribution meets certain smoothness criteria and the log-Sobolev inequality, which essentially confines the applicability of their findings to distributions with a single peak. Recently, Lee et al. (2023) and S. Chen et al. (2023b) have tried to avoid the strong assumptions on the data distributions and to get the polynomial convergence guarantees under general metrics such as TV or Wasserstein distance. Specifically, Lee et al. (2023) give 2-Wasserstein bounds for any distributions with bounded support. Contrary to De Bortoli (2023) and Lee et al. (2022), the results they provide have polynomial complexity guarantees without relying on the functional inequality on distributions such as log-Sobolev inequality. They further give TV bounds with polynomial complexity guarantees under the Hessian availability assumption. Like Lee et al. (2022), under the general data distribution assumption, i.e., the second moment bound of ν and L-Lipschitzness of the score function, S. Chen et al. (2023b) give the polynomial TV convergence guarantee, where only $\tilde{\Theta}\left(\frac{L^2 d}{\varepsilon^2}\right)$ discretization is needed. Here, ε is a TV generalization error, and d is a data dimension.

4.2.2. Attempts to answer question 2. Due to recent theoretical advancements, the list of research attacking the second question is short. M. Chen et al. (2023) proved that the diffusion model is adaptive to estimating the data distribution supported in a lower-dimensional subspace. They design a very specific network architecture for $\mathbf{S}_\theta(\mathbf{x}(t), t)$ with an encoder–decoder structure and a skip connection. Under a more general setting, Oko et al. (2023) prove that the distribution estimator from the diffusion model can achieve nearly minimax optimal estimation rates. Specifically, they assume the true density is supported on $[-1, 1]^d$, in the Besov space with a smooth boundary. The Besov space unifies many general function spaces, such as Hölder, Sobolev, continuous, or even noncontinuous function classes (also refer to Section 2.2). The result of Oko et al. (2023) is valid for the noncontinuous function class, and this should be contrasted with the aforementioned works (Lee et al. 2023, S. Chen et al. 2023b) that assume the Lipschitzness of the score function. The exact architecture of the score network is also given in the form of Equation 2.

4.3. In-Context Learning in Large Language Models

We provide readers with recent theoretical understandings of the interesting ICL phenomenon observed in LLMs. This refers to the ability of LLMs conditioned on a prompt sequence consisting of examples from a task (input–output pairs) along with the new query input to generate the corresponding output accurately. In an example taken from Garg et al. (2022), these models can produce English translations of French words after being prompted on a few such translations, e.g.,

$$\underbrace{\text{maison} \to \text{house}, \quad \text{chat} \to \text{cat}, \quad \text{chien} \to}_{\text{prompt}} \quad \underbrace{\text{dog}}_{\text{completion}} .$$

This capability is quite intriguing as it allows models to adapt to a wide range of downstream tasks on the fly without the need to update the model weights after training. Readers can refer to the backbone architecture of LLMs (Transformer) in the seminal article of Vaswani et al. (2017).

Toward further understanding ICL, researchers formulated a well-defined problem of learning a function class $\mathcal{F}$ from in-context examples. Formally, let $\mathcal{D}_{\mathcal{X}}$ be a distribution over inputs and $\mathcal{D}_{\mathcal{F}}$ be a distribution over functions in $\mathcal{F}$. A prompt P is a sequence $(x_1, f(x_1), \ldots, x_k, f(x_k), x_{\text{query}})$ where inputs (x_i and x_{query}) are drawn independently and identically distributed from $\mathcal{D}_{\mathcal{X}}$ and f is drawn from $\mathcal{D}_{\mathcal{F}}$. In the above example, it can be understood that $\{(x_1, f(x_1), x_2, f(x_2)\} := \{$(maison, house), (chat, cat)$\}$, $x_{\text{query}} =$ chien, and $f(x_{\text{query}}) =$ dog. Now, we provide the formal definition of ICL (Garg et al. 2022).

Definition 2 (In-context learning (ICL)). Model M_θ can in-context learn the function class $\mathcal{F}$ up to ε, with respect to $(\mathcal{D}_{\mathcal{F}}, \mathcal{D}_{\mathcal{X}})$, if it can predict $f(x_{\text{query}})$ with an average error

$$\mathbb{E}_{P\sim(x_1,f(x_1),\ldots,x_k,f(x_k),x_{\text{query}})}\left[\ell(M_\theta(P), f(x_{\text{query}}))\right] \le \varepsilon, \qquad 26.$$

where $\ell(\cdot,\cdot)$ is some appropriate loss function, such as the squared error.

Garg et al. (2022) empirically investigated the ICL of the Transformer architecture (Vaswani et al. 2017) by training the model M_θ on random instances from linear functions, two-layer ReLU networks, and decision trees. Specifically, they showed the predictions of Transformers on the prompt P behave similarly to those of ordinary least squares when the models are trained on instances from linear function classes $\mathcal{F}^{\text{Lin}} := \{f \mid f(x) = w^\top x, w \in \mathbb{R}^d\}$ for random weights $w \sim \mathcal{N}(0, I_d)$. A similar phenomenon was observed for the models trained on sparse linear functions as the predictions behave like those of lasso estimators.

These nice observations sparked numerous follow-up theoretical studies of ICL on internal mechanisms (Akyürek et al. 2023, Dai et al. 2023, Von Oswald et al. 2023), expressive power (Akyürek et al. 2023, Giannou et al. 2023) and generalizations (Y. Li et al. 2023). Among them, Akyürek et al. (2023) and Von Oswald et al. (2023) investigated the behavior of Transformers when trained on random instances from $\mathcal{F}^{\text{Lin}}$ and showed the trained Transformers' predictions mimic those of a single step of GD. They further constructed Transformers that implement such an update. Zhang et al. (2023) recently explicitly proved that the model parameters estimated via gradient flow converge to the global minimizer of the nonconvex landscape of population risk in Equation 26 for learning $\mathcal{F}^{\text{Lin}}$. Nonetheless, the results in the article are based on a linear self-attention layer without softmax nonlinearities and simplified parametrizations. Huang et al. (2023) subsequently considered the single head attention with softmax nonlinearity and proved the trained model through GD did indeed in-context learn $\mathcal{F}^{\text{Lin}}$ under highly stylized scenarios (i.e., simplified parameter settings, orthonormal features). Recently, Chen et al. (2024) generalized the setting to the multi-head attention layer with softmax nonlinearity for ICL multi-task linear regression problems.

The important work of Bai et al. (2023) showed the existence of Transformers that can implement a broad class of standard machine learning (ML) algorithms in context, such as least squares, ridge regression, lasso, and GD for two-layer neural networks. This article goes on to demonstrate a remarkable capability of a single Transformer: the ability to dynamically choose different base ICL algorithms for different ICL instances, all without requiring explicit guidance on the correct algorithm to use in the input sequence. This observation is noteworthy as it mirrors the way statisticians select the learning algorithms for inferences on model parameters.

5. CONCLUSIONS AND FUTURE TOPICS

In this article, we reviewed the literature studying neural networks, mainly from statistical viewpoints. In Section 2, we reviewed statistical literature that primarily relies on approximation-theoretic results of neural networks. This framework allows for interesting comparisons between neural networks and classical linear estimators in various function estimation settings. Specifically, neural networks are highly adaptive to functions with special geometric structures, whereas classical linear estimators are not (see **Figure 1**). In Section 3, we reviewed literature studying the statistical guarantees of neural networks trained with gradient-based algorithms. The overparametrization of neural networks impacts the landscape of loss functions, streamlining the mathematical analysis of training dynamics. We discussed training dynamics in two regimes: NTK and MF. Specifically, we introduced some works that studied how networks in the NTK regime can offer statistical guarantees under noisy observations. We also introduced attempts to unify

and go beyond these regimes, explaining the success of networks with finite widths. In Section 4, we reviewed the statistical guarantees of deep generative models (GANs and diffusion models) for estimating the target distributions. Neural networks form the fundamental backbone of both frameworks, enabling the adaptive estimation of distributions with specialized structures. Some statistical works on ICL phenomena observed in LLM are also introduced. However, aside from these topics, several promising avenues have not yet been covered in this article, and we briefly review them now.

5.1. Generative Data Science

In modern ML, data are valuable but often scarce and have been referred to as "the new oil," a metaphor credited to mathematician Clive Humby. With the rise of deep generative models like GANs, diffusion models, and LLMs, synthetic data are rapidly filling the void of real data and finding extensive utility in various downstream applications. For instance, in the medical field, synthetic data have been utilized to improve patients' data privacy and the performance of predictive models for disease diagnosis (Chen et al. 2021). Similarly, structured tabular data are the most common data modality that requires the employment of synthetic data to resolve privacy issues (X. Li et al. 2023, Suh et al. 2023) or missing values (Ouyang et al. 2023b). Furthermore, synthetic data have been at the core of building reliable AI systems, specifically for the promising fields of fairness (Zeng et al. 2024) and robustness (Ouyang et al. 2023a). Despite the prevalence of synthetic data in the real world, how to evaluate synthetic data from the dimensions of fidelity, utility, and privacy preservation remains unclear. Specifically, we want to address the following general questions: (*a*) How well do the models trained via synthetic data generalize to real unseen data (e.g., Xu et al. 2024)? (*b*) How do artificially generated data perform in the various downstream tasks, such as classification or regression (e.g., X. Li et al. 2023, Xu et al. 2023) and adversarial training (e.g., Xing et al. 2022a,b)? (*c*) How do the synthetic data generated work to satisfy certain privacy constraints (e.g., differential privacy) (Dwork 2008)?

Addressing these questions systematically requires establishing a new framework of generative data science, aiming to elucidate the underlying principles behind generative AI. As evidenced by the above referenced works, this vision is supported by the recent observation that creating something out of nothing is possible and beneficial through synthetic data generation.

5.2. Kolmogorov–Arnold Networks

As of June 2024, a new type of architecture, the Kolmogorov–Arnold network (KAN) (Liu et al. 2024), has been receiving enormous attention from the ML community. The model is motivated by the Kolmogorov–Arnold representation theorem (KART) (Hecht-Nielsen 1987), which states that any continuous and smooth functions on the bounded domain can be represented as compositions and summations of the finite number of univariate functions. Several papers have explored the connections between neural networks and the KART due to their similarities in terms of compositional structure. For instance, readers are directed to Poggio & Girosi (1989), Girosi & Poggio (1989), Schmidt-Hieber (2021), and references therein. Liu et al. (2024) claim that KAN outperforms fully connected networks in terms of accuracy and interpretability for function approximations on their specially designed tasks. Nonetheless, this model definitely requires further research for better use in the future for both practical and theoretical purposes. From a practical perspective, KAN's training time is 10 times slower than fully connected networks, and the authors only apply the model to small tasks (interested readers can refer to Liu et al. 2024, section 6). From a theoretical point of view, they claim that KAN avoids the curse of dimensionality for function approximation, whereas fully connected networks cannot. But this argument requires further investigation under more rigorous settings with various types of function spaces $\mathcal{G}$.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank an anonymous reviewer and Minshuo Chen for insightful comments on the draft of this review. This survey is partially sponsored by the National Science Foundation Stimulating Collaborative Advances Leveraging Expertise in the Mathematical and Scientific Foundations of Deep Learning (NSF–SCALE MoDL) 2134209, National Science Foundation Division of Computer and Network Systems (NS—CNS) 2247795, Office of Naval Research (ONR) N00014-22-1-2680, and a CISCO Research Grant.

LITERATURE CITED

Akyürek E, Schuurmans D, Andreas J, Ma T, Zhou D. 2023. What learning algorithm is in-context learning? Investigations with linear models. arXiv:2211.15661v3 [cs.LG]

Allen-Zhu Z, Li Y. 2019. What can ResNet learn efficiently, going beyond kernels? In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 9017–28. Red Hook, NY: Curran

Allen-Zhu Z, Li Y, Liang Y. 2019a. Learning and generalization in overparameterized neural networks, going beyond two layers. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 6158–69. Red Hook, NY: Curran

Allen-Zhu Z, Li Y, Song Z. 2019b. A convergence theory for deep learning via over-parameterization. arXiv:1811.03962 [cs.LG]

Anderson BD. 1982. Reverse-time diffusion equation models. *Stochast. Process. Appl.* 12(3):313–26

Arora S, Du SS, Hu W, Li Z, Salakhutdinov RR, Wang R. 2019a. On exact computation with an infinitely wide neural net. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 8141–50. Red Hook, NY: Curran

Arora S, Du SS, Hu W, Li Z, Wang R. 2019b. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *Proc. Mach. Learn. Res.* 97:322–32

Arora S, Ge R, Liang Y, Ma T, Zhang Y. 2017. Generalization and equilibrium in generative adversarial nets (GANs). *Proc. Mach. Learn. Res.* 70:224–32

Ba J, Erdogdu MA, Suzuki T, Wang Z, Wu D. 2024. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. In *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, ed. A Oh, T Naumann, A Globerson, K Saenko, M Hardt, S Levine, pp. 17420–49. Red Hook, NY: Curran

Ba J, Erdogdu MA, Suzuki T, Wang Z, Wu D, Yang G. 2022. High-dimensional asymptotics of feature learning: how one gradient step improves the representation. In *NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, ed. S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, A Oh, pp. 37932–46. Red Hook, NY: Curran

Bai Y, Chen F, Wang H, Xiong C, Mei S. 2023. Transformers as statisticians: provable in-context learning with in-context algorithm selection. arXiv:2306.04637 [cs.LG]

Bai Y, Lee JD. 2020. *Beyond linearization: on quadratic and higher-order approximation of wide neural networks.* Paper presented at the International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, Apr. 30

Bai Y, Ma T, Risteski A. 2019. Approximability of discriminators implies diversity in GANs. arXiv:1806.10586 [cs.LG]

Bakry D, Gentil I, Ledoux M. 2014. *Analysis and Geometry of Markov Diffusion Operators*. New York: Springer

Barron AR. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* 39(3):930–45

Barron AR. 1994. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* 14:115–33

Bartlett PL, Harvey N, Liaw C, Mehrabian A. 2019. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.* 20(1):2285–301

Bartlett PL, Montanari A, Rakhlin A. 2021. Deep learning: a statistical viewpoint. *Acta Numer.* 30:87–201

Bauer B, Kohler M. 2019. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Stat.* 47(4):2261–85

Belkin M. 2021. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numer.* 30:203–48

Bengio Y, Roux N, Vincent P, Delalleau O, Marcotte P. 2005. Convex neural networks. In *NIPS'05: Proceedings of the 18th International Conference on Neural Information Processing Systems*, ed. Y Weiss, B Schlkopf, JC Platt, pp. 123–30. Red Hook, NY: Curran

Bietti A, Mairal J. 2019. On the inductive bias of neural tangent kernels. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 12893–904. Red Hook, NY: Curran

Blanchard M, Bennouna MA. 2022. *Shallow and deep networks are near-optimal approximators of Korobov functions*. Paper presented at the International Conference on Learning Representations (ICLR 2022), virtual, Apr. 25

Block A, Mroueh Y, Rakhlin A. 2020. Generative modeling with denoising auto-encoders and Langevin sampling. arXiv:2002.00107 [stat.ML]

Cao Y, Fang Z, Wu Y, Zhou DX, Gu Q. 2020. Towards understanding the spectral bias of deep learning. arXiv:1912.01198 [cs.LG]

Cao Y, Gu Q. 2019. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 10836–46. Red Hook, NY: Curran

Chen M, Huang K, Zhao T, Wang M. 2023. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. arXiv:2302.07194 [cs.LG]

Chen M, Jiang H, Liao W, Zhao T. 2022a. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: function approximation and statistical recovery. *Inf. Inference* 11(4):1203–53

Chen M, Liao W, Zha H, Zhao T. 2022b. Distribution approximation and statistical estimation guarantees of generative adversarial networks. arXiv:2002.03938 [cs.LG]

Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. 2021. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* 5(6):493–97

Chen S, Chewi S, Lee H, Li Y, Lu J, Salim A. 2023a. The probability flow ODE is provably fast. arXiv:2305.11798 [cs.LG]

Chen S, Chewi S, Li J, Li Y, Salim A, Zhang AR. 2023b. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. arXiv:2209.11215 [cs.LG]

Chen S, Sheen H, Wang T, Yang Z. 2024. Training dynamics of multi-head softmax attention for in-context learning: emergence, convergence, and optimality. arXiv:2402.19442 [cs.LG]

Chen Z, Cao Y, Gu Q, Zhang T. 2020. A generalized neural tangent kernel analysis for two-layer neural networks. In *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H. Lin, pp. 13363–73. Red Hook, NY: Curran

Chizat L, Bach F. 2018. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ed. S Bengio, HM Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, pp. 3040–50. Red Hook, NY: Curran

Cybenko G. 1989. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* 2(4):303–14

Dai D, Sun Y, Dong L, Hao Y, Sui Z, Wei F. 2023. Why can GPT learn in-context? Language models secretly perform gradient descent as meta optimizers. arXiv:2212.10559 [cs.CL]

De Bortoli V. 2023. Convergence of denoising diffusion models under the manifold hypothesis. arXiv:2208.05314 [stat.ML]

De Bortoli V, Thornton J, Heng J, Doucet A. 2021. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *NIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 17695–709. Red Hook, NY: Curran

Delalleau O, Bengio Y. 2011. Shallow versus deep sum-product networks. In *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*, ed. J Shawe-Taylor, RS Zemel, PL Bartlett, F Pereira, KQ Weinberger, pp. 666–74. Red Hook, NY: Curran

Devlin J, Chang MW, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL]

DeVore R, Hanin B, Petrova G. 2021. Neural network approximation. *Acta Numer.* 30:327–444

Dhariwal P, Nichol A. 2021. Diffusion models beat GANs on image synthesis. In *NIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 8780–94. Red Hook, NY: Curran

Dong Q, Li L, Dai D, Zheng C, Wu Z, et al. 2024. A survey on in-context learning. arXiv:2301.00234 [cs.CL]

Donoho DL, Johnstone IM. 1998. Minimax estimation via wavelet shrinkage. *Ann. Stat.* 26(3):879–921

Du SS, Lee J, Li H, Wang L, Zhai X. 2019. Gradient descent finds global minima of deep neural networks. *Proc. Mach. Learn. Res.* 97:1675–85

Du SS, Zhai X, Poczos B, Singh A. 2018. *Gradient descent provably optimizes over-parameterized neural networks*. Paper presented at the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, May 6–9

Dwork C. 2008. Differential privacy: a survey of results. In *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008, Proceedings*, ed. M Agrawal, D Du, Z Duan, A Li, pp. 1–19. New York: Springer

Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, et al. 2019. A guide to deep learning in healthcare. *Nat. Med.* 25(1):24–29

Fan J, Ma C, Zhong Y. 2021. A selective overview of deep learning. *Stat. Sci.* 36(2):264–90

Fang Z, Cheng G. 2023. Optimal learning rates of deep convolutional neural networks: additive ridge functions. arXiv:2202.12119 [cs.LG]

Farrell MH, Liang T, Misra S. 2021. Deep neural networks for estimation and inference. *Econometrica* 89(1):181–213

Garg S, Tsipras D, Liang PS, Valiant G. 2022. What can transformers learn in-context? A case study of simple function classes. In *NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, ed. S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, A Oh, pp. 30583–98. Red Hook, NY: Curran

Ghorbani B, Mei S, Misiakiewicz T, Montanari A. 2020. Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function." *Ann. Stat.* 48(4):1898–901

Ghorbani B, Mei S, Misiakiewicz T, Montanari A. 2021a. When do neural networks outperform kernel methods? arXiv:2006.13409 [stat.ML]

Ghorbani B, Mei S, Misiakiewicz T, Montanari A. 2021b. Linearized two-layers neural networks in high dimension. *Ann. Stat.* 49(2):1029–54

Giannou A, Rajput S, Sohn J-y, Lee K, Lee JD, Papailiopoulos D. 2023. Looped transformers as programmable computers. arXiv:2301.13196 [cs.LG]

Girosi F, Poggio T. 1989. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Comput.* 1(4):465–69

Girshick R, Donahue J, Darrell T, Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR '14: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–87. Piscataway, NJ: IEEE

Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial nets. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, ed.

Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 2672–80. Red Hook, NY: Curran
Grigorescu S, Trasnea B, Cocias T, Macesanu G. 2020. A survey of deep learning techniques for autonomous driving. *J. Field Robot.* 37(3):362–86
Gui J, Sun Z, Wen Y, Tao D, Ye J. 2021. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* 35(4):3313–32
Gunasekar S, Lee J, Soudry D, Srebro N. 2018. Characterizing implicit bias in terms of optimization geometry. *Proc. Mach. Learn. Res.* 80:1832–41
Han S, Mao H, Dally WJ. 2016. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv:1510.00149 [cs.CV]
Han S, Pool J, Tran J, Dally W. 2015. Learning both weights and connections for efficient neural network. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, ed. C Cortes, DD Lee, M Sugiyama, R Garnett, pp. 1135–43. Red Hook, NY: Curran
Han Z, Yu S, Lin SB, Zhou DX. 2022. Depth selection for deep ReLU nets in feature extraction and generalization. *IEEE Trans. Pattern Anal. Mach. Intel.* 44(4):1853–68
Hanin B. 2019. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics* 7(10):992
Hayakawa S, Suzuki T. 2020. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Netw.* 123:343–61
He F, Tao D. 2021. Recent advances in deep learning theory. arXiv:2012.10931 [cs.LG]
Heaton JB, Polson NG, Witte JH. 2017. Deep learning for finance: deep portfolios. *Appl. Stochast. Models Bus. Ind.* 33(1):3–12
Hecht-Nielsen R. 1987. Kolmogorov's mapping neural network existence theorem. In *Proceedings of the International Conference on Neural Networks*, Vol. 3, pp. 11–14. New York: IEEE
Ho J, Jain A, Abbeel P. 2020. Denoising diffusion probabilistic models. In *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H. Lin, pp. 6840–51. Red Hook, NY: Curran
Hornik K, Stinchcombe M, White H. 1989. Multilayer feedforward networks are universal approximators. *Neural Netw.* 2(5):359–66
Hornik K, Stinchcombe M, White H. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Netw.* 3(5):551–60
Hu T, Shang Z, Cheng G. 2020. Optimal rate of convergence for deep neural network classifiers under the teacher-student setting. arXiv:2001.06892 [stat.ML]
Hu T, Wang W, Lin C, Cheng G. 2021. Regularization matters: a nonparametric perspective on overparametrized neural network. *Proc. Mach. Learn. Res.* 130:829–37
Hu W, Li Z, Yu D. 2019. *Simple and effective regularization methods for training on noisily labeled data with generalization guarantee*. Paper presented at the International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, Apr. 30
Hu W, Xiao L, Adlam B, Pennington J. 2020. The surprising simplicity of the early-time learning dynamics of neural networks. In *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H. Lin, pp. 17116–28. Red Hook, NY: Curran
Huang Y, Cheng Y, Liang Y. 2023. In-context convergence of transformers. arXiv:2310.05249 [cs.LG]
Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv:1602.07360 [cs.CV]
Imaizumi M, Fukumizu K. 2019. Deep neural networks learn non-smooth functions effectively. *Proc. Mach. Learn. Res.* 89:869–78
Imaizumi M, Fukumizu K. 2022. Advantage of deep neural networks for estimating functions with singularity on hypersurfaces. *J. Mach. Learn. Res.* 23:4772–825
Jacot A, Hongler C, Gabriel F. 2018. Neural tangent kernel: convergence and generalization in neural networks. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ed. S Bengio, HM Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, pp. 8580–89. Red Hook, NY: Curran

Ji Z, Telgarsky M. 2019. Risk and parameter convergence of logistic regression. arXiv:1803.07300 [cs.LG]
Jiang H, Chen Z, Chen M, Liu F, Wang D, Zhao T. 2018. *On computation and generalization of generative adversarial networks under spectrum control*. Paper presented at the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, May 6–9
Jiao Y, Shen G, Lin Y, Huang J. 2021. Deep nonparametric regression on approximately low-dimensional manifolds. arXiv:2104.06708 [math.ST]
Johnson J. 2018. Deep, skinny neural networks are not universal approximators. arXiv:1810.00393 [cs.LG]
Kidger P, Lyons T. 2020. Universal approximation with deep narrow networks. *Proc. Mach. Learn. Res.* 125:2306–27
Kim J, Lee C, Park N. 2022. *STASY: score-based tabular data synthesis*. Paper presented at the International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, May 1–5
Kim Y, Ohn I, Kim D. 2021. Fast convergence rates of deep neural networks for classification. *Neural Netw.* 138:179–97
Kohler M, Krzyżak A, Langer S. 2022. Estimation of a function of low local dimensionality by deep neural networks. *IEEE Trans. Inform. Theory* 68(6):4032–42
Kohler M, Langer S. 2020. Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function." *Ann. Stat.* 48(4):1906–10
Kohler M, Langer S. 2021. On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Stat.* 49(4):2231–49
Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, ed. F Pereira, CJC Burges, L Bottou, KQ Weinberger, pp. 1097–105. Red Hook, NY: Curran
Kutyniok G. 2020. Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function." *Ann. Stat.* 48(4):1902–5
Lai J, Xu M, Chen R, Lin Q. 2023. Generalization ability of wide neural networks on ℝ. arXiv:2302.05933 [stat.ML]
Lee H, Lu J, Tan Y. 2022. Convergence for score-based generative modeling with polynomial complexity. In *NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, ed. S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, A Oh, pp. 22870–82. Red Hook, NY: Curran
Lee H, Lu J, Tan Y. 2023. Convergence of score-based generative modeling for general data distributions. *Proc. Mach. Learn. Res.* 201:946–85
Lee J, Bahri Y, Novak R, Schoenholz SS, Pennington J, Sohl-Dickstein J. 2018. *Deep neural networks as Gaussian processes*. Paper presented at the International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Can., Apr. 30–May 3
Li X, Wang C, Cheng G. 2023. Statistical theory of differentially private marginal-based data synthesis algorithms. arXiv:2301.08844 [cs.LG]
Li Y, Ildiz ME, Papailiopoulos D, Oymak S. 2023. Transformers as algorithms: generalization and implicit model selection in in-context learning. arXiv:2301.07067 [cs.LG]
Li Y, Liang Y. 2018. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ed. S Bengio, HM Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, pp. 8168–77. Red Hook, NY: Curran
Li Z, Wang R, Yu D, Du SS, Hu W, et al. 2019. Enhanced convolutional neural tangent kernels. arXiv:1911.00809 [cs.LG]
Liang T. 2021. How well generative adversarial networks learn distributions. *J. Mach. Learn. Res.* 22(1):10366–406
Liu X, Wu L, Ye M, Liu Q. 2022. Let us build bridges: understanding and extending diffusion generative models. arXiv:2208.14699 [cs.LG]
Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, et al. 2024. KAN: Kolmogorov-Arnold networks. arXiv:2404.19756 [cs.LG]
Lu J, Shen Z, Yang H, Zhang S. 2021. Deep network approximation for smooth functions. *SIAM J. Math. Anal.* 53(5):5465–506

Lu Z, Pu H, Wang F, Hu Z, Wang L. 2017. The expressive power of neural networks: a view from the width. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, ed. U von Luxburg, I Guyon, S Bengio, H Wallach, R Fergus, pp. 6232–40. Red Hook, NY: Curran

Mei S, Bai Y, Montanari A. 2018a. The landscape of empirical risk for nonconvex losses. *Ann. Stat.* 46(6A):2747–74

Mei S, Misiakiewicz T, Montanari A. 2019. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *Proc. Mach. Learn. Res.* 99:2388–464

Mei S, Montanari A, Nguyen PM. 2018b. A mean field view of the landscape of two-layer neural networks. *PNAS* 115(33):E7665–71

Mhaskar H, Liao Q, Poggio T. 2017. When and why are deep networks better than shallow ones? In *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 2343–49. Washington, DC: AAAI

Mhaskar HN. 1996. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.* 8(1):164–77

Montanelli H, Du Q. 2019. New error bounds for deep ReLU networks using sparse grids. *SIAM J. Math. Data Sci.* 1(1):78–92

Montufar GF, Pascanu R, Cho K, Bengio Y. 2014. On the number of linear regions of deep neural networks. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 2924–32. Red Hook, NY: Curran

Müller A. 1997. Integral probability metrics and their generating classes of functions. *Adv. Appl. Probability* 29(2):429–43

Müller-Franzes G, Niehues JM, Khader F, Arasteh ST, Haarburger C, et al. 2022. Diffusion probabilistic models beat GANs on medical images. arXiv:2212.07501 [eess.IV]

Nakada R, Imaizumi M. 2020. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.* 21(174):1–38

Neelakantan A, Vilnis L, Le QV, Sutskever I, Kaiser L, et al. 2015. Adding gradient noise improves learning for very deep networks. arXiv:1511.06807 [stat.ML]

Neyshabur B. 2017. Implicit regularization in deep learning. arXiv:1709.01953 [cs.LG]

Nguyen PM, Pham HT. 2023. A rigorous framework for the mean field limit of multilayer neural networks. arXiv:2001.11443 [cs.LG]

Nitanda A, Suzuki T. 2020. *Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime*. Paper presented at International Conference on Learning Representations (ICLR 2021), Vienna, Austria, May 4

Novak R, Xiao L, Lee J, Bahri Y, Yang G, et al. 2020. Bayesian deep convolutional networks with many channels are Gaussian processes. arXiv:1810.05148 [stat.ML]

Oko K, Akiyama S, Suzuki T. 2023. Diffusion models are minimax optimal distribution estimators. arXiv:2303.01861 [stat.ML]

Otter DW, Medina JR, Kalita JK. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* 32(2):604–24

Ouyang Y, Xie L, Cheng G. 2023a. Improving adversarial robustness through the contrastive-guided diffusion process. *Proc. Mach. Learn. Res.* 202:26699–723

Ouyang Y, Xie L, Li C, Cheng G. 2023b. MissDiff: training diffusion models on tabular data with missing values. arXiv:2307.00467 [cs.LG]

Oymak S, Soltanolkotabi M. 2020. Toward moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE J. Sel. Areas Inform. Theory* 1(1):84–105

Park S, Yun C, Lee J, Shin J. 2020. Minimum width for universal approximation. arXiv:2006.08859 [cs.LG]

Pascanu R, Montufar G, Bengio Y. 2014. On the number of response regions of deep feed forward networks with piece-wise linear activations. arXiv:1312.6098 [cs.LG]

Petersen P, Voigtlaender F. 2018. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.* 108:296–330

Poggio T, Girosi F. 1989. *A theory of networks for approximation and learning.* Rep. AIM 1140, Artif. Intell. Lab., Cambridge, MA

Poole B, Lahiri S, Raghu M, Sohl-Dickstein J, Ganguli S. 2016. Exponential expressivity in deep neural networks through transient chaos. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, ed. DD Lee, U von Luxburg, R Garnett, M Sugiyama, I Guyon, pp. 3368–76. Red Hook, NY: Curran

Roweis ST, Saul LK. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–26

Ryzhik L. 2023. *Lecture notes for Math 272, winter 2023*. Lecture Notes, Dep. Math., Stanford Univ., Stanford, CA

Särkkä S, Solin A. 2019. *Applied Stochastic Differential Equations*. Cambridge, UK: Cambridge Univ. Press

Schmidt-Hieber J. 2020. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Stat.* 48(4):1875–97

Schmidt-Hieber J. 2021. The Kolmogorov–Arnold representation theorem revisited. *Neural Netw.* 137:119–26

Schreuder N, Brunel VE, Dalalyan A. 2021. Statistical guarantees for generative models without domination. *Proc. Mach. Learn. Res.* 132:1051–71

Shamir O. 2020. Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function." *Ann. Stat.* 48(4):1911–15

Shen Z, Yang H, Zhang S. 2021. Deep network approximation characterized by number of neurons. arXiv:1906.05497 [math.NA]

Smith S, Elsen E, De S. 2020. On the generalization benefit of noise in stochastic gradient descent. *Proc. Mach. Learn. Res.* 119:9058–67

Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. 2020. *Score-based generative modeling through stochastic differential equations*. Paper presented at International Conference on Learning Representations (ICLR 2021), Vienna, Austria, May 4

Song Z, Yang X. 2020. Quadratic suffices for over-parametrization via matrix Chernoff bound. arXiv:1906.03593 [cs.LG]

Suh N, Ko H, Huo X. 2021. *A non-parametric regression viewpoint: Generalization of overparametrized deep ReLU network under noisy observations*. Paper presented at the International Conference on Learning Representations (ICLR 2022), virtual, Apr. 25

Suh N, Lin X, Hsieh DY, Honarkhah M, Cheng G. 2023. AutoDiff: combining auto-encoder and diffusion model for tabular data synthesizing. arXiv:2310.15479 [stat.ML]

Suh N, Zhou TY, Huo X. 2022. *Approximation and non-parametric estimation of functions over high-dimensional spheres via deep ReLU networks*. Paper presented at the International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, May 1–5

Suzuki T. 2018. *Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality*. Paper presented at the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, May 6–9

Suzuki T, Nitanda A. 2021. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *NIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, ed. M Ranzato, A Beygelzimer, Y Dauphin, PS Liang, J Wortman Vaughan, pp. 3609–21. Red Hook, NY: Curran

Tang R, Yang Y. 2022. Minimax rate of distribution estimation on unknown submanifold under adversarial losses. arXiv:2202.09030 [math.ST]

Tenenbaum JB, de Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–23

Van de Geer SA. 2000. *Empirical Processes in M-Estimation*. Cambridge, UK: Cambridge Univ. Press

Vardi G, Yehudai G, Shamir O. 2022. Width is less important than depth in ReLU neural networks. *Proc. Mach. Learn. Res.* 178:1249–81

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, ed. U von Luxburg, I Guyon, S Bengio, H Wallach, R Fergus, pp. 6000–10. Red Hook, NY: Curran

Von Oswald J, Niklasson E, Randazzo E, Sacramento J, Mordvintsev A, et al. 2023. Transformers learn in-context by gradient descent. *Proc. Mach. Learn. Res.* 202:35151–74

Wainwright MJ. 2019. *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge, UK: Cambridge Univ. Press

Wei C, Lee JD, Liu Q, Ma T. 2019. Regularization matters: generalization and optimization of neural nets versus their induced kernel. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 9712–24. Red Hook, NY: Curran

Wu X, Du SS, Ward R. 2019. Global convergence of adaptive gradient methods for an over-parameterized neural network. arXiv:1902.07111 [cs.LG]

Xing Y, Song Q, Cheng G. 2022a. Unlabeled data help: minimax analysis and adversarial robustness. *Proc. Mach. Learn. Res.* 151:136–68

Xing Y, Song Q, Cheng G. 2022b. Why do artificially generated data help adversarial robustness? In *NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, ed. S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, A Oh, pp. 954–66. Red Hook, NY: Curran

Xu S, Sun WW, Cheng G. 2024. Utility theory of synthetic data generation. arXiv:2305.10015 [stat.ML]

Xu S, Wang C, Sun WW, Cheng G. 2023. Binary classification under local label differential privacy using randomized response mechanisms. *Trans. Mach. Learn. Res.* **https://openreview.net/forum?id=uKCGOw9bGG**

Yang G, Hu EJ. 2022. Feature learning in infinite-width neural networks. arXiv:2011.14522 [cs.LG]

Yang L, Zhang Z, Song Y, Hong S, Xu R, et al. 2024. Diffusion models: a comprehensive survey of methods and applications. arXiv:2209.00796 [cs.LG]

Yarotsky D. 2017. Error bounds for approximations with deep ReLU networks. *Neural Netw.* 94:103–14

Yehudai G, Shamir O. 2019. On the power and limitations of random features for understanding neural networks. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 6598–608. Red Hook, NY: Curran

Zeng X, Dobriban E, Cheng G. 2024. Bayes-optimal classifiers under group fairness. arXiv:2202.09724 [stat.ML]

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64(3):107–15

Zhang P, Liu Q, Zhou D, Xu T, He X. 2018. On the discrimination-generalization tradeoff in GANs. arXiv:1711.02771 [cs.LG]

Zhang R, Frei S, Bartlett PL. 2023. Trained transformers learn linear models in-context. arXiv:2306.09927 [stat.ML]

Zhong G, Wang LN, Ling X, Dong J. 2016. An overview on data representation learning: from traditional feature learning to recent deep learning. *J. Finance Data Sci.* 2(4):265–78

Zou D, Cao Y, Zhou D, Gu Q. 2018. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. arXiv:1811.08888 [cs.LG]

Zou D, Gu Q. 2019. An improved analysis of training over-parameterized deep neural networks. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, pp. 2055–64. Red Hook, NY: Curran