A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions: A Review

Wenhao Zhong

2100010609
School of Mathematical Sciences
Peking Univeristy

Yang Lv

2000010793
School of Mathematical Sciences
Peking University

Bofan Chen

1900011030
School of Economics
Peking University

Abstract

We first introduce the background of the research and list some works related to the research. We then give a brief summary of the content of the paper. This is followed by examples on MNIST to exemplify the three IPMs suggested by the paper and we discuss the significance of the paper. Finally, we give remarks on the limitations of and potential directions inspired by the research.

1 Introduction

1.1 Background

Recent years, deep learning has achieved state-of-the-art performance in lots of applications, such as image recognition, machine translation, game-playing. Mathematically speaking, deep neural networks can be thought as a function mapping from input x to the output y, where x is the data and y can be a vector, probability distribution and so on. It is like polynomial, but performs much better and more efficient in high dimensions than polynomial. It is natural to believe that the deep neural networks' structure has good ability to approximate mapping in high dimensions. However, the advantages of deep neural networks are far from understood. Despite the successful applications in reality, the theory of the neural networks is limited. We don't know why it can work so well and efficiently in high dimensions. What's more, many models are not robust and many deep learning projects meet the bottleneck in development due to the lack of rigor theory. Therefore, it is necessary and meaningful to understand the neural network structure using mathematics to get some theoretical basis.

1.2 Related works

It is interesting to study the universal approximation theorem and the expressive power of a deep learning model. There are lots of research trying to understand the neural network structure. One of the most classic results is the "Universal Approximation Theorem", which indicates that any continuous functions can be approximated by a two-layer neural network. However, it doesn't show why the neural networks are efficient. Recently, many researchers tend to investigate how depth affects the expressive capacity of a neural network. There are some developments in understanding approximating functions. In particular, with a deep ReLU neural network, [1] proved that for n-variate continuous functions, we can control the uniform error by its amplitude and the number of the parameters. Dmitry Yarotsky [5] analyzed the high-order continuity situation, where we can control the uniform error with certain depth and weights, which indicates that deep networks can approximate smooth functions than shallow ones more efficiently. [3] also discussed the advantage of

Preprint. Under review.

deep networks over shallow ones in classification problems. [2] analyzed n-variate piecewise smooth functions to control its L^2 approximation error. [4] discussed the wide neural networks, shows that there exist some wide networks can't be approximated by networks whose depth is no more than a polynomial bound, which can let us have a better understanding of the width and depth. These are all theorems about approximating the functions, and some of the results are optimal.

Compared with approximating the functions, there are fewer results for approximating distributions. Goodfellow et al. [9] proved that Generative Adversarial Nets models can converge to the data distribution if the model is big enough and the algorithm is good enough. However, like the "Universal Approximation Theorem", it also doesn't show why the GAN models are efficient. The expressive power of generative networks has been studied by Lee et al.[6], who showed that for those distributions mapped by a series of Barron function from a noisy distribution can be approximated by a neural network to control the Wasserstein distance. However, we don't know exactly what distribution can be mapped by Barron functions, and the methods used by them can't apply to the case of distributions which are not under Barron functions mapping. [7] analyzed the conversion of two uniform distributions in different dimensions and the conversion of a uniform distribution and a normal distribution in the same dimension. But these are just simple distributions and they didn't analyze those more general distributions. Another important result related to this topic but not strongly correlated to this work is that the expressive power of normalizing flow models is limited in high dimensions [8].

2 Summary

This paper mainly focuses on the question that how well DNNs can express probability distributions.

To better define the problem, The authors use three different integral probability metrics (IPMs) to measure the closeness between probability distributions: 1-Wasserstein metric($\mathcal{W}_1$), maximum mean discrepancy(MMD), and kernelized Stein discrepancy(KSD). The three IPMs are defined as follows:

$$\mathcal{W}_1(p,\pi) = \inf_{\gamma\in\Gamma(p,\pi)} \int |x-y|\gamma(dxdy)$$

$$\mathrm{MMD}(p,\pi) = \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \left|\mathbf{E}_{X\sim p} f(X) - \mathbf{E}_{X\sim\pi} f(X)\right|.$$

$$\mathrm{KSD}(p,\pi) = \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \mathbf{E}_{X\sim p}\left[\nabla\log\pi\cdot f(X) + \nabla\cdot f(X)\right]$$

Given a desired approximation error ε under three IPMs, the paper aims at deriving the upper bounds for the depth and width of the DNN.

The overall proof idea can be mainly divided into two parts.

In first part, the authors consider DNNs to approximate the discrete probability distribution. This part makes use of the conclusions of semi-discrete optimal transport maps. It is obvious that any discrete distribution $\nu = \sum_{j=1}^n \nu_j\delta_{y_j}$ can be derived from the source distribution μ (i.e. standard normal) through transformations $T:\mathbb{R}^d\to\mathbb{R}^d$, namely $\nu = T_{\#}\mu$. Under the background of semi-discrete optimal transport, we need to find a optimal transport map T that minimize the quadratic cost as follows.

$$\inf_T \int \frac{1}{2}|x-T(x)|^2\mu(dx)\ s.t \int_{T^{-1}(y_j)} d\mu = \nu_j, j = 1,\cdots,n$$

This problem can be simplified by maximizing the following functional

$$\mathcal{F}(\psi) = \sum_{j=1}^n \left[\int_{P_j}\left(\frac{1}{2}\left|x-y_j\right|^2 - \psi_j\right)\rho(x)dx + \psi_j\nu_j\right]$$

where

$$P_j := \left\{x\in\mathbb{R}^d | \frac{1}{2}\left|x-y_j\right|^2 - \psi_j \leq \frac{1}{2}\left|x-y_k\right|^2 - \psi_k, \forall k\neq j\right\}$$

The optimal transport T is given by $T(x) = \nabla\bar{\varphi}(x)$ where $\bar{\varphi}(x) = \frac{1}{2}|x|^2 - \varphi(x)$ and $\min_j\left\{\frac{1}{2}\left|x-y_j\right|^2 - \psi_j\right\}$. Surprisingly, we can finally derive $\bar{\varphi}(x) = \max_j\left\{x\cdot y_j + m_j\right\}$ with

$m_j = \psi_j - \frac{1}{2}\left|y_j\right|^2$. This kind of form means that T can be represented by the gradient of a DNN: Just let $n = 2^k$ and consider

$$\bar{\varphi}(x) = \max_{j=1,\cdots,2^k}\{x \cdot y_j + m_j\} = \max_{j=1,\cdots,2^{k-1}} \max_{i\in\{2j-1,2j\}}\{x \cdot y_i + m_i\}$$

and

$$\max(a, b) = \operatorname{ReLU}(a - b) + \operatorname{ReLU}(b) - \operatorname{ReLU}(-b)$$

So for any discrete probability distribution ν, we can represent $\nu = (\nabla\bar{\varphi})_{\#}\mu$ where $\bar{\varphi}(x)$ is a DNN of depth $L = \lceil \log n \rceil$ and width $N = 2^L = 2^{\lceil \log n \rceil}$ and $\mu \in \mathcal{P}_2\left(\mathbb{R}^d\right)$ is the source distribution.

In the second part, the authors consider a special discrete distribution, namely the empirical distribution P_n of the target distribution π. We have already known that we can represent the discrete empirical distribution P_n with the gradient of a DNN. So the problem becomes what the distance $D(P_n, \pi)$ is between π and P_n under three IPMs respectively. Obviously, we may get different P_n for the same π. Our problem requires us to show the existance of P_n that is really close to π. The authors transform the problem of existence into a problem of probability. That is to say, if we show that $D(P_n, \pi)$ decreases in a certain rate under a postive probability, we will complete the proof.

For the three given IPMs, the authors derive the rate of approximation as follows:

- If π satisfies $M_3 = \mathbf{E}_{X\sim\pi}|X|^3 < \infty$, then there exists a realization of empirical measure P_n such that
$$\mathcal{W}_1\left(P_n, \pi\right) \leq C \cdot \begin{cases} n^{-1/2}, & d = 1 \\ n^{-1/2}\log n, & d = 2 \\ n^{-1/d}, & d \geq 3 \end{cases}$$
- If $\sup_{x\in\mathbb{R}^d}|k(x, x)| \leq K_0$, then there exists a realization of empirical measure P_n such that
$$\operatorname{MMD}\left(P_n, \pi\right) \leq \frac{C}{\sqrt{n}},$$
- If π is L-Lipschitz and sub-Gaussian, and k is twice differentiable which satisfies $\max_{m+n\leq 1}\sup_{x,y}\left\|\nabla_x^m\nabla_y^n k(x, y)\right\| \leq K_1$ and $\sup_{x,y}\left|\operatorname{Tr}\left(\nabla_x\nabla_y k(x, y)\right)\right| \leq K_1(1 + d)$, then there exists a realization of empirical measure P_n such that
$$\operatorname{KSD}\left(P_n, \pi\right) \leq C\sqrt{\frac{d}{n}},$$

The derivation of the KSD is worth mentioning, and it is one of the main contributions of the authors in this paper.

Combining the conclusions of both parts, the authors derive that given the distance $D(P_n, \pi) < \varepsilon$ the approximate width n of the corresponding DNN is as follows: (the approximate depth is $\lceil \log n \rceil$)

- $D = \mathcal{W}_1$, then $n \leq \begin{cases} \frac{C}{\varepsilon^2}, & d = 1 \\ \frac{C\log^2(\varepsilon)}{\varepsilon^2}, & d = 2 \\ \frac{C^d}{\varepsilon^d}, & d \geq 3 \end{cases}$
- $D = \operatorname{MMD}$, then $n \leq \frac{C}{\varepsilon^2}$
- $D = \operatorname{KSD}$, then $n \leq \frac{Cd}{\varepsilon^2}$

This conclusion means that the complexity bound may be suffered from Curse of Dimention(CoD) when $D = \mathcal{W}_1$, whereas the complexity bound is not suffered from Curse of Dimention when $D = \operatorname{MMD}$ and $D = \operatorname{KSD}$.

3 Examples

The main theorem shows that for different IPMs, the upper bound of the networks has a huge difference. Wasserstein Distance suffers from CoD, while the result of Maximum Mean Discrepancy, unexpectedly, does not relate to the dimension. It is strange because the model should be bigger as the dimension increases, so we doubt whether Maximum Mean Discrepancy can reflect the discrepancy of two distributions well. We trained original GAN and Wasserstein GAN on the MNIST digits dataset (the Generator is with the same network structure, fully connected 28*28, 128, 256, 512, 1024, 28*28 units, ReLU, RMSProp, the only difference is the loss functions). The following figure presents the digits learned after 200 iterations. For MMD GAN (the structure is proposed by Gintare et al.), we tried many times but didn't get good results. The following figure is from Gintare et al.'s paper [10, Training generative neural networks via Maximum Mean Discrepancy optimization]. They trained 1,000,000 iterations (it needs too much time, it may be a reason we didn't get good result, we trained 2000 iterations and it already took more than 5 hours). However, even they trained 5000 times than ours, the result seems not very good compared with GAN and Wasserstein GAN. Though it is possible that the algorithm proposed by Gintare et al. is not as good as GAN, it is also highly possible that the Maximum Mean Discrepancy can't measure the difference of two distributions well in practice. Also, in practice, it is hard to compute these distances directly. Wasserstein GAN used a method to approximate Wasserstein Distance, but according to our experiments, the result of original GAN (with BCELoss as loss function) is better than Wasserstein GAN on MNIST dataset. If we want to use these IPMs as loss functions, how to design an algorithm to approximate them is another problem.

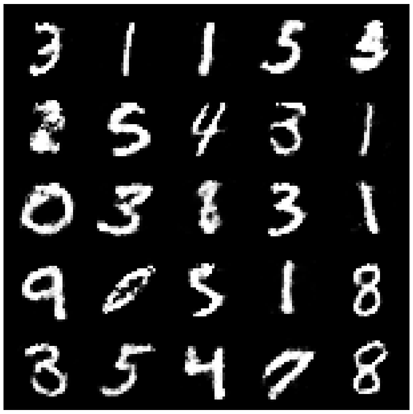

Figure 1: GAN

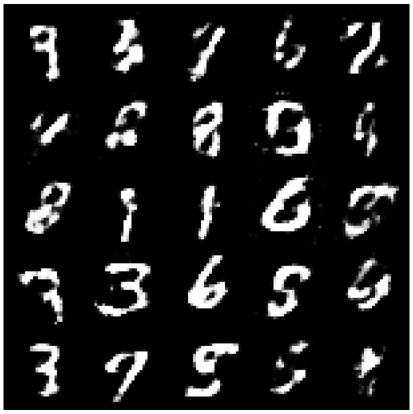

Figure 2: Wasserstein GAN

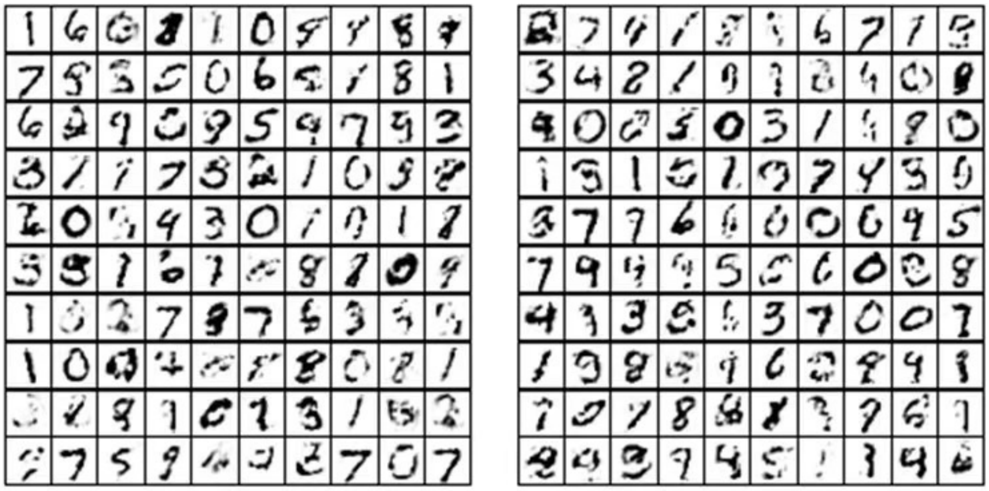

Figure 3: GAN with Maximum Mean Discrepancy optimization

4 Significance

The paper is significant because it is the first paper to raise the question of and give the explicit answer to the universal approximation properties of DNNs. This may inspire following works to work on analysis of specific generative models, or other more in-depth properties of DNNs for expressing probability distributions. These questions are crucial for us to understand the performance of modern deep generative models, which is seldom addressed by existng literature. The writing of this paper is also clear, elegant and intuitive, which may bring audiences outside of the deep learning theory community to appreciate and work on the result.

5 Limitations and potential research directions

Here, we remark this paper from some perspectives.

The CoD problem The authors argue that the use of distance $\mathcal{W}_1$ leads to CoD while the use of the other two distances does not lead to CoD. I think this is a rather strange way to look at CoD, and it feels like we are cheating ourselves by using another metric to avoid CoD, while the probability distribution itself remains unchanged. In fact, discrete distributions themselves like the empirical distributions are intuitively not good approximations for the continues target distribution. Under MMD and KSD, the empirical distribution, however, does not have the problem of CoD. Does this mean that the two metrics themselves have some defects? Shouldn't we reflect on the problem of metric itself instead of continuing to cover our ears? Recalling the discussion of the approximation rate of functions by neural networks, it is clear that we cannot make an approximation to function in any function space and avoid CoD at the same time. Our solution is to restrict the space of the objective function, and then approximate them to avoid CoD. I think it is good to transfer this idea to the problem of approximating probability distributions, we do not need to consider our target function π into the full function space, but fix it in some common function space in daily life. For example, we can restrict the continuity of the function, the moment condition, some condition of its Fourier transform, *etc.*, then approximate and conclude that there is no CoD in this function space under a reasonable distance metric. Methodologically, I think it is also possible to use Monte Carlo methods as a basis for deriving the space of functions for which our desired probability distribution is satisfied.

Problems arising from discrete distributions By reading the authors' proof, we find that the authors use the empirical distribution (discrete distribution) as a bridge to approximate. Intuitively, the discrete distribution is more difficult to obtain directly with a DNN (continuous) . Also, we learned from the knowledge of KDE that using a discrete distribution $\nu = \sum_{j=1}^{n} \nu_j \delta_{y_j}$ to approximate an probability distribution is not cost-effective. Therefore, using the empirical distribution as a bridge to approximate the target probability distribution may not be a good choice, although there are some existing conclusions in optimal transport problem we can use. Perhaps, in the future, we can try to use a continuous distribution as a bridge instead of a discrete distribution. For example, we can assume that our target distribution is continues, and we use $\nu(y) = \sum_{j=1}^{n} \nu_j K(\frac{y-y_j}{h})$ as a bridge of proof, where $K(\cdot)$ is kernel function.

Transformation between different dimensions In this paper, the authors consider only the transformation $T : \mathbb{R}^D \to \mathbb{R}^d$ between equal dimensions (*i.e.*$D = d$). however, in practice, there are many cases where we need to consider transformations between different dimensions(*i.e.*$D \neq d$). Obviously, if $D < d$, we cannot approximate a probability function with d-dimensions. In the case of $D > d$, the upper bound of the neural network have not discussed, which may be a good potential research direction.

Problems with using gradient of DNN to pushforward distribution Despite some real world generative models indeed use gradient of a DNN to pushforward the source distribution to target distribution as mentioned in the paper, the majority of models don't do it, e.g. classical GAN, variational autoencoder, normalizing flow and diffusion models. The authors justify their approach in the paper as "Because DNNs are continuous but the map from source to target distribution is not continuous in practice.". We argue here that this cannot fully explain the case, as discontinuous maps can always be smoothed with a small perbutation. If you successfully learn the smoothed map, you can generate high quality images with very low probability of generating a wrong sample. Think of the case of supervised learning for image classification: the target function is discrete but a continuous DNN addresses this problem very well.

Remarks on technical aspect We now turn ourselves to the technical(proof) aspect of the paper. Checking through the paper, it seems that the authors throws a combination of well-known results, and didn't make enough new theoretical analysis. To be more specific, the authors first described the convergence behavior of 3 empirical measures, with 2 of them already known before. The authors then discussed semi-discrete optimal transport with solution of a piecewise affine function, which is also well known in optimal transport literature. Finally, the authors proved that DNN gradients can approximate piecewise linear affine functions, which has been well understood by the deep learning community before. It seems that in order to reach their conclusion, the authors just picked some known results, thus limiting the use of their theorem (and perhaps making the result not deep enough).

Explaining real world performance Universal approximation property is only the first step to understand DNNs for expressing probablistic distributions. It may be questionable that this paper, like many other papers on deep learning theory, can actually explain why real world deep generative models prefer well. The setting in this paper seems only directly applies to some variants of GANs. It can help us understand the universal approximation properties of certain variants of GANs with loss function designed as certain metric, but this doesn't necessary explain the performance of these GANs in practice. It is fairly possible that instead of approximation error, optimization error can be the actual bottleneck of the applications of these GANs. Also, the paper provides some upper bounds of the width and depth of DNNs needed to obtain certain approximation error. This seems not that valuable, especially those on the bounds of length due to the reason that today the state-of-the-art performance of generative modelling task is often obtained by very large models with billions of nodes and we can train very deep neural networks using the technique of skip connections(residue networks).

Beyond simple DNN structures It is important to notice that the main result of this paper is a existence result which states that, in short, certain DNNs can approximate the target distribution well under certain metrics. However, it says nothing about deep generative models that is not the same architecture as the paper suggests. As said in previous subsection, the result of this paper can only possibly be appiled to certain variants of GANs. Although this is certainly beyond the scope of this paper, it is interesting and important to consider the problem of explaining performance of other generative models, especially those reaching or once reached a state-of-the-art performance. There has been seldom discussion of these issues in the literature and we will be providing some thoughts here.

Optimal transport theory is a crucial part of the proof the theorem. Optimal tranport theory can be used to understand many deep generative models. We know that using optimal transport, we can derive [11] a weak version of variational autoencoder(VAE) and it is probable that one can give a theoretical analysis of the universal approximation property of VAE similar to the one given in this paper. In recent years diffusion models have reached state-of-the-art of generative modelling. The diffusion models provide striking success in tasks like text-to-image generation, image super-resolution, colorization, etc. Diffusion model learns an invertible distribution between the source distribution and limit of the diffusion distribution. Numerical evidence[12] shows that this map is numerically very close to the Monge optimal transport map. If one wants to analysis the universal approximation properties of diffusion model, he may take adventage of (or try to prove equality of) this fact. But this is not a semi-discrete map and one has to study the convergence of continuous measures to target function.

6 Conclusion

This paper gives a theoretical result about the error using DNN to approximate distribution, but it is still far from understanding the expressive power of deep learning. We proposed some limitations of this paper and potential research directions, hoping to get better theoretical result and have a deeper understanding of how theorem can guide practical applications. We also raised the question about how well different IPMs can reflect the difference of two distributions, which is helpful to guide designing the loss functions in practice.

References

[1] Dmitry Yarotsky (2018) Optimal approximation of continuous functions by very deep ReLU networks.

[2] Philipp Petersen & Felix Voigtlaender (2017) Optimal approximation of piecewise smooth functions using deep ReLU neural networks

[3] Matus Telgarsky (2015) Representation Benefits of Deep Feedforward Networks.

[4] Zhou Lu & Hongming Pu & Feicheng Wang & Zhiqiang Hu & Liwei Wang (2017) The Expressive Power of Neural Networks: A View from the Width. *31st Conference on Neural Information Processing Systems*

[5] Dmitry Yarotsky (2017) Error bounds for approximations with deep ReLU networks.

[6] Holden Lee & Rong Ge & Tengyu Ma & Andrej Risteski & Sanjeev Arora (2017) On the Ability of Neural Nets to Express Distributions. *Proceedings of Machine Learning Research vol 65:1–26, 2017*

[7] Bolton Bailey & Matus Telgarsky (2018) Size-Noise Tradeoffs in Generative Networks.

[8] Zhifeng Kong & Kamalika Chaudhuri (2020) The Expressive Power of a Class of Normalizing Flow Models. *AISTATS, Volume 108.*

[9] Ian J. Goodfellow et al. (2014) Generative Adversarial Nets.

[10] Gintare Karolina Dziugaite & Daniel M. Roy & Zoubin Ghahramani (2015) Training generative neural networks via Maximum Mean Discrepancy optimization.

[11]Genevay, Aude, Gabriel Peyré, and Marco Cuturi. "GAN and VAE from an optimal transport point of view." arXiv preprint arXiv:1706.01807 (2017).

[12]Khrulkov, Valentin, and Ivan Oseledets. "Understanding ddpm latent codes through optimal transport." arXiv preprint arXiv:2202.07477 (2022).