Literature Review

Rohan Hitchcock

In the last few decades the use of neural networks for machine learning (deep learning) has seen widespread adoption in many commercial and scientific areas. It has been demonstrated in multiple domains (image processing [1], natural language processing [2], game playing [3], [4], protein folding [5]) that this approach is highly capable of solving complex and data intensive problems. Furthermore, experimental evidence suggests that the capability of neural network models will continue to improve as the size of datasets grow [6]–[8].

Despite the technical success of deep learning, many parts of the theory of deep learning are not well understood and classical statistical theories do not apply [9]. Singular learning theory, which is developed by Sumio Watanabe in [10], provides a framework to develop the theory of deep learning.

Singular learning theory

Suppose we are given the task of learning the true probability distribution $q(x)$ of some data by considering a family of probability distributions $\{p(x|w)\}_{w\in W}$, which we call a statistical model. For example, each $p(x|w)$ is a neural network with weights $w \in W$ and we wish to find $w^* \in W$ such that $q(x) = p(x|w^*)$. An important theoretical tool is the *Kullback-Leibler distance* $K : W \to [0, \infty]$ [10, Definition 1.1 p. 3], which is defined for $w \in W$ as

$$K(w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx.$$

The Kullback-Leibler distance can be thought of as measuring how far the approximating distribution $p(x|w)$ is from the true distribution $q(x)$. It can be shown that $K(w) = 0$ if and only if $p(x|w) = q(x)$, and so we define the set of true parameters as $W_0 = \{w \in W : K(w) = 0\}$.

A statistical model is *regular* if the map $w \mapsto p(x|w)$ is injective and the Kullback-Leibler distance can be approximated by a quadratic form when w is close to the true parameter (assuming $W_0 \neq \emptyset$) [10, Definition 1.7, Remark 1.4 pp. 8, 10]. Singular learning theory studies statistical models which are not necessarily regular, and neural networks are one type of type of non-regular statistical model [10, p. 11].

Blow ups and resolution of singularities

A central tool in singular learning theory is Hironaka's Resolution of Singularities [11] which uses the concept of a 'blow up' from algebraic geometry to simplify the singularities of algebraic and analytic sets.

Let k be a field. In the application of the following to singular learning theory we will have $k = \mathbb{R}$. Following Chapter 1 of [12] we define an *affine algebraic set* as a subset of k^n which is the zero set of a collection of polynomial functions from $k[x_1, \cdots, x_n]$, and an *affine algebraic variety* as an algebraic set which is not the union of two other non-empty algebraic sets. We also define *projective n-space* as $\mathbf{P}^n = k^{n+1} \setminus \{0\} / \sim$,

where $\boldsymbol{x} \sim \boldsymbol{y}$ if and only if $\boldsymbol{x} = \lambda \boldsymbol{y}$ for some $\lambda \in k$, or equivalently as the set of lines passing through the origin of k^{n+1}. The *blow up* [12, p. 28] of k^n at $0 \in k^n$ is defined as $X = \{(\boldsymbol{x}, \boldsymbol{y}) \in k^n \times \mathbf{P}^{n-1} : x_i y_j - y_j y_i = 0,\ i, j = 1, \cdots, n\}$. Let $\varphi : X \to k^n$ denote the usual projection restricted to X. One can show that $X \setminus \varphi^{-1}(0) \simeq k^n \setminus \{0\}$ and $\varphi^{-1}(0) \simeq \mathbf{P}^{n-1}$, so, informally speaking, the blow up of k^n replaces the origin $0 \in k^n$ with a copy of $\mathbf{P}^{n-1}$. The blow up of a variety $V \subseteq k^n$ is found by considering the preimage $\varphi^{-1}(V)$. If a variety V has a singularity at 0, then blowing up the variety has the effect of removing the singularity. Each point of $\varphi^{-1}(0)$ corresponds to a line passing through the origin, and it can be shown that a curve passing through 0 in k^n will intersect $\varphi^{-1}(0)$ in X at a point corresponding to the slope of the curve at 0. As we have defined it here, the definition of the blow up of an affine variety depends on its embedding in k^n, however from the modern, scheme-theoretic point of view of algebraic geometry the blow up is defined via a universal property [12, Chapter 2.7], and so is revealed as a more fundamental operation on varieties.

Stated for polynomial functions, Hironaka's Resolution of Singularities (of [11], and stated in Theorem 3.5, Theorem 3.6 of [10, pp. 97, 98]), says that by using repeated blow ups we can obtain a paramterisation of any affine algebraic variety in which the singularities are of a particular, simple form called normal crossing form. There is also an version for analytic functions. We define an *analytic set* to be the zero set of finitely many analytic functions. Stated for analytic functions Hironaka's Resolution of Singularities says that for any analytic set there exists a local reparameterisation of every singularity that puts it in normal crossing form [11], [10, Theorem 2.3 p. 58], and furthermore these local reparameterisations can be combined in a compatible way [13], [10, Remark 3.6 p. 8].

Resolution of singularities in singular learning theory

Hironaka's Resolution of Singularities arises in singular learning theory because the Kullback-Leibler distance $K(w)$ is an analytic function of w, and so the set of true parameters of a statistical model $W_0 = \{w \in W : K(w) = 0\}$ is an analytic set. Applying the Resolution of Singularities can help us prove theorems about singular models like neural networks. Given the Kullback-Leibler distance of a model $K(w)$ and a prior probability density $\psi(w)$ on the parameter space W we can define the *zeta function* [10, p. 217] of the model as

$$\zeta(z) = \int K(w)^z \psi(w) dw.$$

The poles of $\zeta(z)$ encode important information about the statistical model. For example, it can be shown that the Bayes generalisation error of the model can be expressed in terms of the maximal pole and order $\zeta(z)$ [14]. In [15] and [10, Example 7.1 p. 225] the poles of $\zeta(z)$ are computed explicitly for a particular, simple neural network by finding the Resolution of Singularities reparameterision.

More generally singular learning theory – which describes models in terms of things like the poles of $\zeta(z)$ – can be used to explain apparent contradictions between the predictions of regular learning theory and the observed behaviour of neural networks [9]. It is hoped that the approach to analysing singular models developed in [10] can be used to make progress on currently open problems in the theory of deep learning.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 24, 2017, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386. [Online]. Available: https://dl.acm.org/doi/10.1145/3065386 (visited on 04/14/2021).

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv:2005.14165 [cs]*, Jul. 22, 2020. arXiv: 2005.14165. [Online]. Available: http://arxiv.org/abs/2005.14165 (visited on 04/14/2021).

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, Number: 7587 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature16961. [Online]. Available: https://www.nature.com/articles/nature16961 (visited on 04/14/2021).

[4] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature24270. [Online]. Available: http://www.nature.com/articles/nature24270 (visited on 04/14/2021).

[5] E. Callaway, "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures," *Nature*, vol. 588, no. 7837, pp. 203–204, Nov. 30, 2020, Number: 7837 Publisher: Nature Publishing Group. DOI: 10.1038/d41586-020-03348-4. [Online]. Available: https://www.nature.com/articles/d41586-020-03348-4 (visited on 04/14/2021).

[6] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, C. Hallacy, B. Mann, A. Radford, A. Ramesh, N. Ryder, D. M. Ziegler, J. Schulman, D. Amodei, and S. McCandlish, "Scaling laws for autoregressive generative modeling," *arXiv:2010.14701 [cs]*, Nov. 5, 2020. arXiv: 2010.14701. [Online]. Available: http://arxiv.org/abs/2010.14701 (visited on 04/14/2021).

[7] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv:1712.00409 [cs, stat]*, Dec. 1, 2017. arXiv: 1712.00409. [Online]. Available: http://arxiv.org/abs/1712.00409 (visited on 04/14/2021).

[8] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv:2001.08361 [cs, stat]*, Jan. 22, 2020. arXiv: 2001.08361. [Online]. Available: http://arxiv.org/abs/2001.08361 (visited on 04/14/2021).

[9] D. Murfet, S. Wei, M. Gong, H. Li, J. Gell-Redman, and T. Quella, "Deep learning is singular, and that's good," *arXiv:2010.11560 [cs]*, Oct. 22, 2020. arXiv: 2010.11560. [Online]. Available: http://arxiv.org/abs/2010.11560 (visited on 04/15/2021).

[10] S. Watanabe, *Algebraic geometry and statistical learning theory*. Cambridge ; New York: Cambridge University Press, 2009, 286 pp., ISBN: 978-0-521-86467-1.

[11] H. Hironaka, "Resolution of singularities of an algebraic variety over a field of characteristic zero," *Annals of Mathematics*, pp. 109–326, 1964, Publisher: JSTOR.

[12] R. Hartshorne, *Algebraic Geometry*, ser. Graduate Texts in Mathematics. New York, NY: Springer New York, 1977, vol. 52, ISBN: 978-1-4419-2807-8 978-1-4757-3849-0. DOI: 10.1007/978-1-4757-3849-0. [Online]. Available: http://link.springer.com/10.1007/978-1-4757-3849-0 (visited on 08/05/2020).

[13] M. F. Atiyah, "Resolution of singularities and division of distributions," *Communications on Pure and Applied Mathematics*, vol. 23, no. 2, pp. 145–150, Mar. 1970, ISSN: 00103640, 10970312. DOI: 10.1002/cpa.3160230202. [Online]. Available: http://doi.wiley.com/10.1002/cpa.3160230202 (visited on 03/22/2021).

[14] S. Watanabe, "Algebraic analysis for nonidentifiable learning machines," *Neural Computation*, vol. 13, no. 4, pp. 899–933, 2001, Publisher: MIT Press.

[15] M. Aoyagi and S. Watanabe, "Resolution of singularities and the generalization error with bayesian estimation for layered neural network," *IEICE Trans*, pp. 2112–2124, 2005.