Singular learning theory VI

sdl6
①
10/6/20

There are a growing number of remarkable experimental findings in deep learning, for which it would be desirable to have a theoretical explanation. These include:

- The generalisation puzzle: the fact that deep learning works at all, and moreover continues to improve over many orders of magnitude in dataset size, model size and computational resources, is already striking. See (sdl) for a more precise discussion.

- Power laws: as observed in [H], [OA1], [OA2] there is experimental evidence in the context of Transformer models trained on language tasks for the following power law [OA1, (1.2)]

$$L(n) = \left(n_c/n\right)^{0.095} \qquad n_c \sim 5.4 \times 10^{13} \qquad (1.1)$$

where $L(n)$ is the test loss for a large model trained with early stopping on a dataset of size n.

What light can singular learning theory shed on these findings? Since neural networks are singular the theory is arguably necessary but in its current state it is not sufficient. In this note we address the key obstacles to applying singular learning theory to modern deep neural networks:

I) The predictive distribution seems irrelevant to deep learning practice
II) ReLU networks are not analytic
III) The true distribution is not realisable
IV) Estimating the learning coefficient at scale is impractical

I. The predictive distribution seems irrelevant

We consider a compact space W of neural network weights and model class $p(y|x,w)$ as in (Forw). The Bayesian posterior associated to a dataset D_n (i.e. a training set of size n sampled from the true distribution $q(y|x)q(x)$) is derived as follows [W, §1.3.1] from Bayes' rule:

$$p(w|D_n) = \frac{p(D_n|w)p(w)}{p(D_n)}$$

$$= \frac{1}{Z_n}\varphi(w)\prod_{i=1}^{n} p(Y_i|X_i,w) \qquad (2.1).$$

where $D_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, $\varphi(w)$ is the prior and Z_n is a normalising constant. The predictive distribution p^* is

$$p^*(y|x) = p(y|x,D_n) = \int p(y|x,w)\,p(w|D_n)\,dw \qquad (2.2)$$

Singularity learning theory is largely concerned with the predictive distribution, as for example the central quantity in the theory is the Bayesian generalisation error (cf. (1.1) of (sdl5))

$$B_g(n) := D_{KL}(q \,\|\, p^*)$$

$$= \int q(y|x)\,q(x)\log\frac{q(y|x)}{p^*(y|x)}\,dx\,dy \qquad (2.3)$$

This is clearly not what people mean by generalisation error in deep learning: they mean the loss on the test set of a single model $p(y|x,w^*)$ where w^* is obtained by SGD training. Since we cannot easily sample from the Bayesian posterior we cannot compute the predictive distribution, and so its performance appears to be irrelevant to deep learning practice. So is singular learning theory talking about a quantity $B_g(n)$ that is of only theoretical interest?

Not quite! The response of a statistician might be the following argument, taken from [WBIC]. Note that

$$Z_n = \int \prod_{i=1}^{n} p(y_i \mid x_i, w)\, \varphi(w)\, dw \qquad (3.1)$$

is a function on $S = (\mathbb{R}^{in} \times \mathbb{R}^{out})^n$ where the model inputs $x \in \mathbb{R}^{in}$ and outputs $y \in \mathbb{R}^{out}$, and that (see [W, Remark 1.10])

$$\begin{aligned} \int Z_n\, q(x)\, dx\, dy &= \int \Big(\prod_{i=1}^{n} \int p(y_i \mid x_i, w)\, dy_i \Big)\, q(x)\, \varphi(w)\, dx\, dw \\ &= \int q(x)\, \varphi(w)\, dx\, dw \;=\; 1 \qquad (3.2) \end{aligned}$$

Hence $Z_n : S \longrightarrow \mathbb{R}$, called the <u>evidence</u>, <u>marginal likelihood</u> or <u>partition function</u>, may be used in the following way: given a proposed pair $(p(y|x,w), \varphi(w))$ we prepare the marginal likelihood Z_n and wait for the sample D_n to arrive. If the model is a "good fit" in that Z_n evaluated on that sample is high, then the "evidence" for the pair $(p(y|x,w), \varphi(w))$ is high.

If Z_n is small, then $Z_n q(x)$ as a distribution on S is assigning probability mass to areas that are not realised by the true distribution, and $p(y|x,w), \varphi(w)$ is therefore a bad fit.

As Watanabe writes in the introduction to [WBIC], the <u>Bayes free energy</u>

$$F_n = -\log Z_n \qquad (\text{so } Z_n = e^{-F_n}) \qquad (3.3)$$

is a decreasing function of Z_n, so in statistical model evaluation you want to choose models with <u>high evidence</u> Z_n and hence low <u>free energy</u> F_n. As detailed in [W], [WBIC] the free energy is related to the predictive distribution and $B_g(n)$.

However, here we encounter a gap: nobody forces me to adopt model selection according to the likelihood principle elaborated above. This is a foundational question of statistics. If deep learning practitioners compared neural network architectures by comparing their evidence in the above sense and selecting the one with highest evidence then singular learning theory would enable comparison of architectures by e.g. estimation of RLCTs (since these are related to F_n).

Perhaps practitioners should do this, but the problem is to explain generalisation for deep neural networks on the terms dictated by deep learning practitioners, not on terms dictated from the outside by statistical authorities. So we have a problem: how to relate "performance on the test set" as a metric of model selection to the evidence Z_n?

I.1 From test loss to Gibbs generalisation error

Here is how model selection in deep learning actually works. Let us take computer vision as an example. The ImageNet dataset was published in 2009 at CVPR [I]. For a detailed history see [L]. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) ran from 2010–2017 and it was the 2012 entry AlexNet in this challenge which is widely credited with widespread interest in deep learning models. This represents one of the most historically decisive instances of model selection in practice (people switching from other kinds of models to deep learning).

The ILSVRC worked as follows: every year the organisers published a training set of 1.2 million images (see image-net.org). Contestants train their models on this set D_n, $n = 1.2 \times 10^6$, and the winner was decided according to performance on a held-out test set T_m of $m = 1.5 \times 10^4$ images, sampled from the same true distribution (images collected from Flickr and search engines) as the training set.

In machine learning generalisation error (meaning error on the test set) is the primary method of model selection [GBC, § 5.2], as the example of ILSVRC demonstrates. Going forward I will use the term "generalisation error" in this sense and use "Bayesian generalisation error" to mean $B_g(n)$.

Note further than in practice the field of deep learning is organised internally around datasets such as ImageNet, and competitions such as ILSVRC, with fixed training sets D_n against which multiple contestants train a range of models, say $p_1(y|x,w), \ldots, p_r(y|x,w)$. We could imagine two competition protocols

(A) contestants train their models against D_n some number of times and submit $\{p_i(y|x,w_i^*)\}_{i=1}^r$ to the competition, where w_i^* is the set of weights for the "best" trained model of team i (leaving aside what "best" means). The teams are then ranked according to the performance of $p_i(y|x,w_i^*)$ on the test set.

(B) contestants train their models against D_n some number of times and submit all of their models, so say team i submits weights $\{w_{i,j}^*\}_{j=1}^s$ where s is the number of runs. The teams are then ranked according to their mean test error averaged over their s models.

In practice (A) is what people do in competitions (because (B) seems hard to enforce) and (B) is roughly how good papers work: people report mean and standard deviation of test error over some number of training runs, against a standard dataset.

How does model selection actually work? Suppose that AlexNet topped the 2012 leaderboard, beating non-deep learning models, but that when people went home and tried to train AlexNet themselves they found that its generalisation error was typically worse than their favourite alternative model. Clearly they will not use AlexNet. They will conclude deep learning doesn't "really" work, and that the AlexNet authors simply got incredibly lucky with their random seeds. This procedure is closer to (B), leading us to

Hypothesis I : model selection in deep learning practice ranks models by their mean test set error, over many independent SGD runs, against fixed training sets (e.g. ImageNet, CIFAR).

Subject to this hypothesis, we now relate model selection in deep learning practice to singular learning theory. Fix a class of models $p(y|x,w)$ and prior $\varphi(w)$ (e.g. a prescription for weight regularisation). Suppose we run SGD training s times against a fixed training set D_n and obtain weights $w_1^*, \ldots, w_s^*$. For each weight we compute the test error for a test set T_m

$$\hat{t}_j := \frac{1}{m} \sum_{(y,x) \in T_m} \log q(y|x) / p(y|x, w_j^*) \qquad (6.1)$$

and then compute $\frac{1}{s} \sum_{j=1}^{s} \hat{t}_j$. The best way of comparing models trained on the common dataset D_n would be to compute the true generalisation error

$$t_j := \mathbb{E}_{(Y,X)} \left[\log q(Y|X) / p(Y|X, w_j^*) \right] \qquad (6.2)$$

for the unknown true distribution, to which $\hat{t}_j$ is an empirical estimate. And ideally one would take s as large as possible, so that model selection is, according to Hypothesis I, performed using the following quantity:

$$G_g^{SGD}(n) := \mathbb{E}_\omega^{SGD}\left[\mathbb{E}_{(Y,X)}\left[\log q(Y|X)/p(Y|X,\omega) \right] \right] \qquad (7.1)$$

where the outside expectation is an average over SGD runs, and corresponds to the empirical distribution of endpoints of SGD training over W. This distribution is complex and depends on many factors (initialisation, SGD variant, learning rate schedule, early stopping etc.). We call (7.1) the SGD Gibbs generalisation error and refer to $p^{SGD}(\omega|D_n)$, the probability of ω being an endpoint of SGD training against D_n, as the SGD posterior.

Hypothesis II The SGD posterior equals the Bayesian posterior for some choice of prior $\varphi^{SGD}(\omega)$.

This is almost certainly false as stated, but we do expect the SGD posterior and Bayesian posterior to be related. As a first rough approximation, to get the theory off the ground, Hypothesis II may be useful even if false. It is certainly "less false" than the hypotheses currently underpinning most deep learning theory!

The Gibbs generalisation error as defined in [W, Def^n 1.8] is

$$G_g(n) := \mathbb{E}_\omega\left[\mathbb{E}_{(Y,X)}\left[\log q(Y|X)/p(Y|X,\omega) \right] \right] \qquad (7.2)$$

where the outside expectation is taken with respect to the Bayesian posterior. Both $G_g^{SGD}(n)$ and $G_g(n)$ are random variables, since they are numbers depending on a sampled training set D_n.

Observation I Under hypotheses I, II model selection in deep learning practice is performed via the Gibbs generalisation error $G_g(n)$ (lower is better).

It remains to relate $G_g(n)$ to $B_g(n)$ and the RLCT, in order to complete the connection between the central characters of singular learning theory and deep learning practice. Here is where things start to get really interesting.

If the true distribution is realisable and the other fundamental conditions hold, [W, Thm. 6.8] ensures the existence of random variables B_g^*, G_g^* such that as $n \longrightarrow \infty$ we have convergence in law

$$n B_g(n) \longrightarrow B_g^*, \quad n G_g(n) \longrightarrow G_g^* \tag{8.1}$$

and $\mathbb{E}[\, n B_g(n)\,] \longrightarrow \mathbb{E}[B_g^*]$, $\mathbb{E}[\, n G_g(n)\,] \longrightarrow \mathbb{E}[G_g^*]$, where these expectations are with respect to D_n. By [W, Thm 6.10] we have

$$\mathbb{E}[\, B_g^*\,] = \lambda, \qquad \mathbb{E}[\, G_g^*\,] = \lambda + \nu \tag{8.2}$$

where λ is the learning coefficient (you may think of this as the RLCT) and ν is the singular fluctuation, another birational invariant. This would directly relate the generalisation error $G_g(n)$ used in deep learning practice to the algebro-geometric quantities λ, ν, via the asymptotic approximation as $n \to \infty$

$$\mathbb{E}[\, G_g(n)\,] \approx \frac{\lambda + \nu}{n} \tag{8.3}$$

However in typical deep learning problems on large real-world datasets the true distribution will never be realisable (see § III). This is consistent with the incompatibility between (8.3) which would predict a scaling exponent 1 (i.e. $\mathbb{E}[\, G_g(n)\,] \propto \frac{1}{n}$) and the very strong empirical evidence (1.1) for a scaling exponent (in this specific case) of 0.095.

So what are we to do? What does Watanabe say in the non-realisable case?

First of all we should note another obstacle, which is that ReLU networks are not even analytic; however this seems relatively unimportant, see §II. The key theoretical difficulties in applying singular learning theory appear to be

Open problem I.1 : prove Hypothesis II, or find the right statement

Open problem I.2 : prove that feedforward ReLU networks, convnets and Transformers satisfy the conditions of learnability [WA] for some constant $\gamma > 0$ depending on the model and true distribution.

To explain, note that the book [W] says very little about the non-realisable case except for some scattered remarks, e.g. Remark 8.3 (1) which is backed by the experiments in §8.3.1, which tend to suggest $\mathbb{E}[G_g^*] \approx \mathbb{E}[B_g^*]$ in the non-realisable case. I do not see strong grounds for believing this. Section 7.6 briefly treats some aspects of the non-realisable case, but not the relation between $G_g(n)$ and λ, ν.

The paper [WBIC] works almost entirely in the not necessarily realisable case, and we know in this generality that F_n is related to λ and may be estimated by the WBIC. Further $B_g(n)$ (denoted G in [WBIC]) is related to $\mathbb{E}[\text{WAIC}]$. However the equations of state and "universal laws" required to relate $G_g(n)$ to $B_g(n)$ and the birational invariants λ, ν are not treated there.

The state of the art for non-realisable models appears to be [WA], in which we find general universal laws stated under a weaker condition, called "conditions of learnability" with coefficient γ, under which we expect formulas such as

$$\mathbb{E}[G_g(n)] \approx L_0 + \frac{\mathcal{O}}{n^\gamma} \qquad (9.1)$$

where $\mathcal{O}$ is some constant, presumably related to λ, ν.

If a class of models is renormalisable then the conditions of learnability hold for $\gamma = 1$, so that (1.1) strongly suggests deep learning models are <u>non-renormalisable</u>.

<u>Observation II</u> if singular learning theory is to apply to deep learning models in a useful way, we need to show the universal laws apply, and we expect these models to be <u>singular</u>, <u>non-realisable</u> and <u>non-renormalisable</u> so that we are genuinely in unknown territory from the point of view of Watanabe's existing work.

If we can show e.g. Transformers satisfy the conditions of learnability for γ, then we would have [WA, (18)]

$$\mathbb{E}[G_g(n)] \approx L_0 + F_n'(0) \qquad (10.1)$$

and the expectation seems to be that even in the non-renormalisable case that $F_n'(0) = \mathcal{O}/n^\gamma$ for some $\mathcal{O}$. It remains to characterise the geometric meaning of $\mathcal{O}$. In any case, this yields

$$\underbrace{\log\left(\mathbb{E}[G_g(n)] - L_0\right)}_{\text{log scale test loss}} \approx \log \mathcal{O} - \gamma \log(n) \qquad (10.2)$$

where we expect $\mathcal{O}$ to depend in an intricate way on the geometry of the set of true parameters. The exponent γ is much more mysterious. It certainly depends on the true distribution ($\gamma = 1$ if the true distribution is realisable) but it is unclear how sensitive it is to the aforementioned geometry (noting that in the realisable or renormalisable cases it is perfectly <u>insensitive</u> to this structure)

In short, Open Problem I.2 goes genuinely beyond the limit of today's singular learning theory.

I.2 Evidence for power laws

There is compelling experimental evidence for power laws (10.2) across a range of modalities and architectures, see [H] and [OA1]. In the former paper they found that the scaling exponent σ primarily depended on the dataset and was surprisingly insensitive to changes in architecture. This was also remarked on in [OA1]. If an architecture is big enough to fit the data, and can scale with compute, their hypothesis is that the scaling exponent is relatively independent of "how" you allocate the weights (e.g. increased width vs. depth). This can only be true within a class of relatively similar models (e.g. LSTMs do not have the same exponent as Transformers).

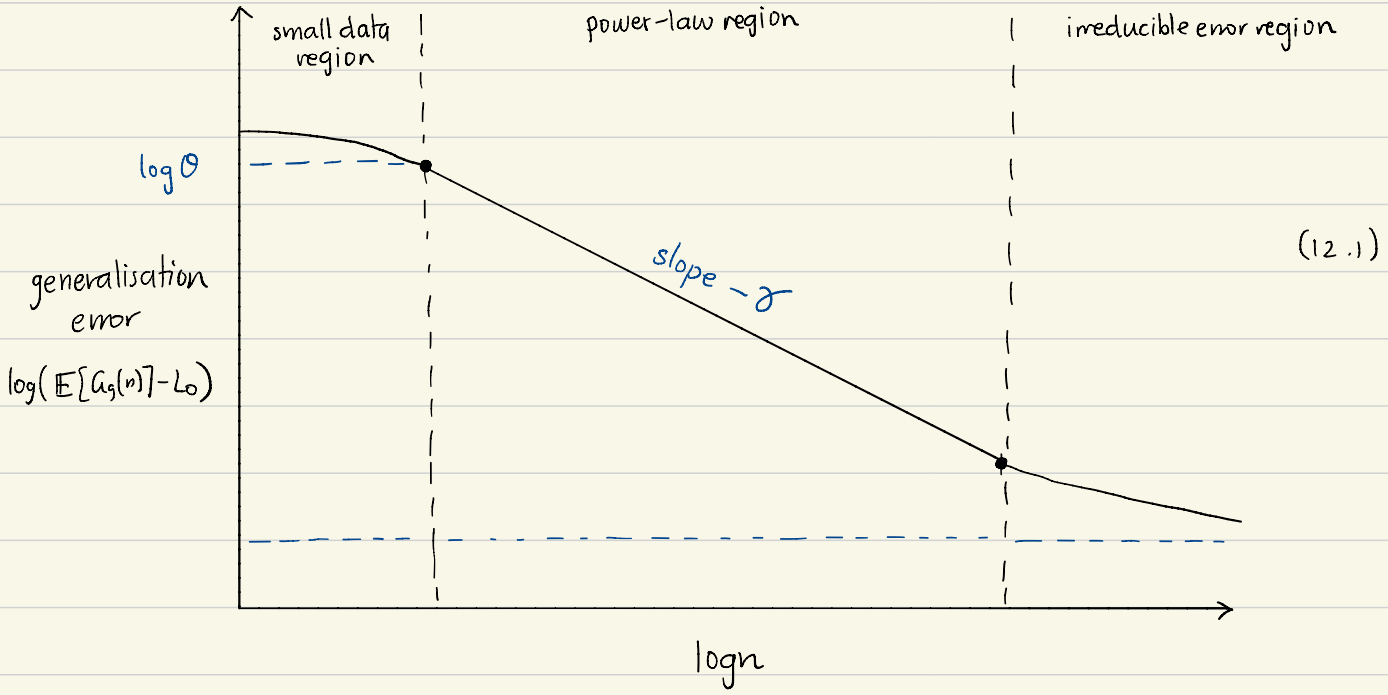
Beating the Power-law: Machine learning researchers often try to improve model accuracy by changing model architectures trained on a given data set. Their efforts can involve complex trial-and-error and rely on creativity or epiphany to improve results. Our tests suggest that model architecture improvements such as model depth only shift learning curves down, but might not improve the power-law exponent.

A broader question is whether machine learning techniques could improve the power-law learning curve exponent, or in other words, to improve generalization more quickly as training data grows. Theory suggests that best case accuracy scaling is with $\beta_p = -0.5$ or -1. Thus, for some problem domains—especially language modeling—the potential accuracy improvements are immense if we knew ways to improve the power-law exponent.

We have yet to find factors that affect the power-law exponent. To beat the power-law as we increase data set size, models would need to learn more concepts with successively less data. In other words, models must successively extract more marginal information from each additional training sample. This might be difficult without adjustments to the data set. We suggest that future work more deeply analyze learning curves when using data handling techniques, such as data filtering/augmentation, few-shot learning, experience replay, and generative adversarial networks.

This is from §5.2 of [H]. In short, they endorse the view that ML researchers are working hard to change Θ in (10.2) but all that ultimately matters (see Sutton's "Bitter lesson") is σ

At this point we return to Watanabe's assertion in the preface to [W] that "knowledge to be discovered from examples corresponds to a singularity". Consider this in the context of the following graph [H, Fig. 6] which is consistent with (10.2) as an asymptotic approximation as $n \to \infty$



(12.1)

Recall that in the realisable case $\theta = \lambda + \nu$ and $\gamma = 1$. Putting aside ν which is hard to understand, the more singular the model (λ small) the lower the starting point of the power-law line. But this ultimately less important than the slope γ.

Consider two models obeying power laws (10.2) with parameters (θ, γ), $(\theta + \delta\theta, \gamma + \delta\gamma)$ and let $\log n$ be large enough that power-law scaling applies to both. The crossover point is the solution of

$$\log\theta - \gamma \log n = \log(\theta + \delta\theta) - (\gamma + \delta\gamma)\log n$$

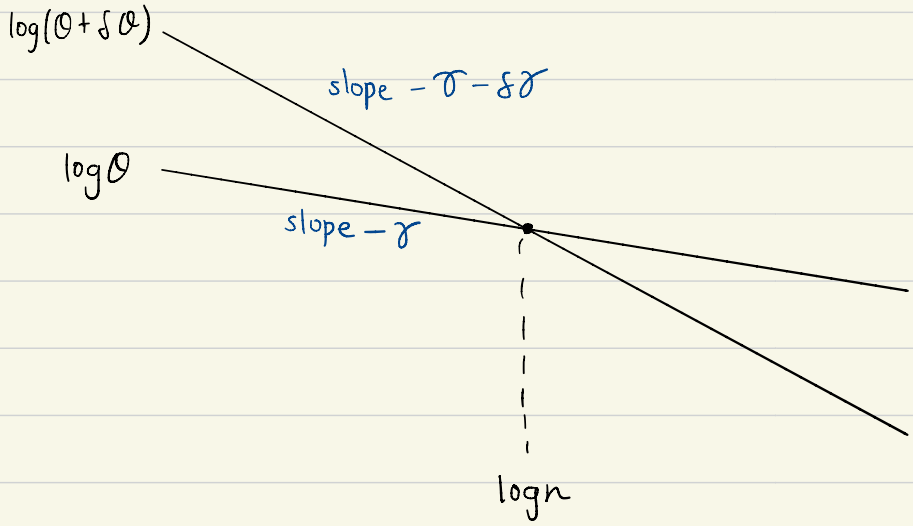
$$\Rightarrow \quad \delta\gamma \log n = \log\left(1 + \frac{\delta\theta}{\theta}\right) \qquad (12.2)$$

$$\therefore \quad \log n = \frac{1}{\delta\gamma}\log\left(1 + \frac{\delta\theta}{\theta}\right) \approx \frac{\delta\theta}{\theta \cdot \delta\gamma} \quad \text{for } \delta\theta/\theta \text{ small.}$$

Suppose $\delta\theta$ is positive (so that the model $\theta+\delta\theta$ looks, say from the point of view of WBIC, to have lower evidence recalling that WBIC ~ free energy so higher WBIC means lower evidence) and let $\log n$ be large enough to be in the power scaling regime. For the model $(\theta+\delta\theta, \gamma+\delta\gamma)$ to beat (θ, γ) at $\log n$ we need

$$\delta\gamma \geq \frac{\delta\theta}{\theta \cdot \log n} \qquad (13.1)$$

which for large data ($\log n \gg 0$) is very small relative to $\delta\theta$ (unless $\theta \ll 0$, and since this related to the "effective number of parameters" in the true distribution it does seem bounded below and not many orders of magnitude less than 1)



(13.2)

So the model (θ, γ) has higher evidence by the standards of regular statistics or the WBIC, but since it "learns less" from each new sample it is quickly surpassed in the large data regime by the model $(\theta+\delta\theta, \gamma+\delta\gamma)$. This seems to represent an even more profound challenge to the conventional wisdom of statistical learning theory even than Watanabe's work, since it suggests that in the large data, large compute regime model selection should be performed based on the scaling exponent γ (larger is better), not according to the free energy.

To quote Ilya Sutskever (chief scientist of OpenAI), on deep learning in Podcast #94 with Lex Fridman

61:03 "The Transformer is the most important advance in neural network architectures in recent history

61:32 "The Transformer is successful because it is the simultaneous combination of multiple ideas and if you were to remove either idea it would be much less successful. So the Transformer uses a lot of attention but attention existed for a few years... The Transformer is designed in such a way that it runs really fast on the GPU, and that makes a huge amount of difference. The second thing is the Transformer is not recurrent and that is really important too because it is more shallow and thus easier to optimise."

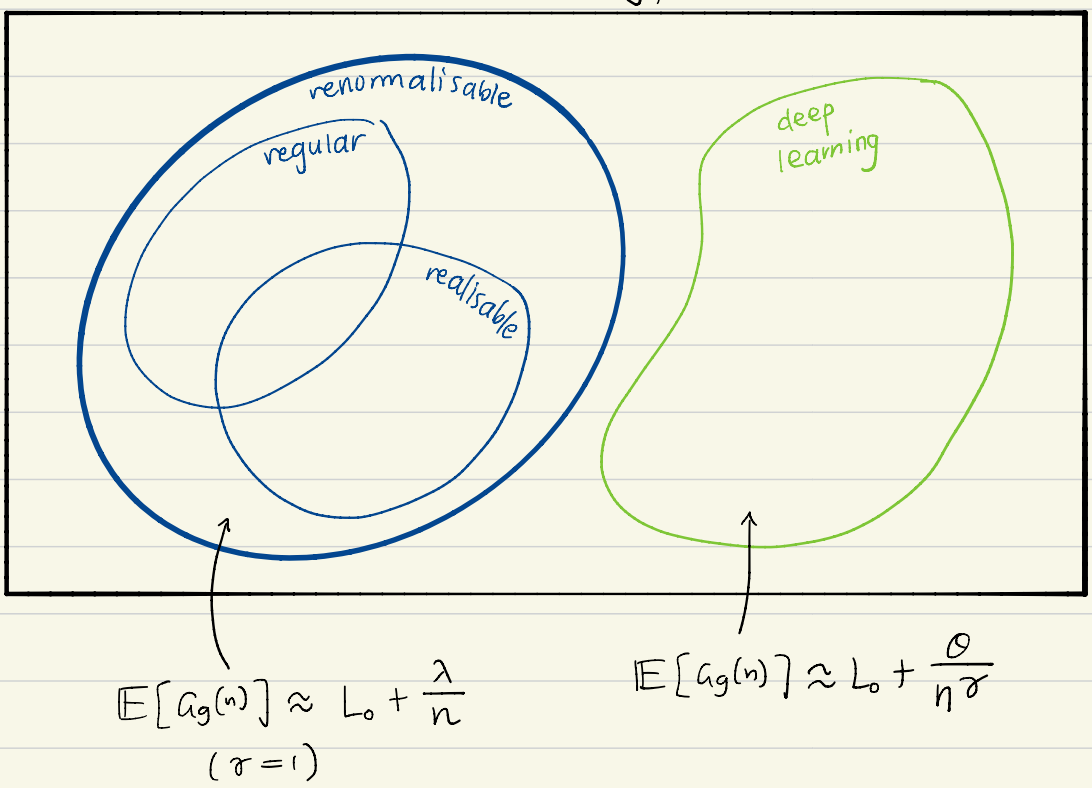
In short, Transformers:

(a) Use attention (architecture)
(b) Are a really great fit for a GPU (scaling with compute)
(c) Not recurrent, so easier to optimise (optimisation)

This is clearly a modern instance of model selection (Transformers > LSTMs) but only (a) is conceivably about θ or the WBIC. Clearly (b) is about "getting on the power law train" and (c) is about dynamics of actually being able to find a good model within the class.

So wither the free energy? In the deep learning era does the WBIC or RLCT even matter, except insofar as they relate to scaling exponents, computational scaling or ease of optimisation?

statistical learning problems



In conclusion, deep learning represents a new phase in statistical learning theory. It is organised around classes of models that are neither regular, nor realisable, or renormalisable, and the operational methods of model selection at scale are arguably very different from the prevailing statistical paradigm.

Observation III We do not care about estimating the learning coefficient of deep learning models, because these numbers do not determine model selection. However we do care about their theoretical existence, and singular learning theory, because it is the best hope of proving the existence of power laws and investigating indices of learnability γ.

Note that phenomena like the gradient noise scale and other dynamical properties of SGD training may be sensitive to the characteristics of $\mathbb{F}$ and thus the geometry. But if the WBIC is not a principal method of model selection it does tend to undermine the argument that algebraic geometry should be central to deep learning practice, except insofar as it is necessary to formulate and prove the underlying theory.

III. The true distribution is not realisable

If $K \subseteq \mathbb{R}^n$ is compact then feedforward ReLU networks of width $m+n$ and arbitrary depth are dense in the space of continuous functions $K \longrightarrow \mathbb{R}^m$ [HS]. Arguably then the true distribution is realised by <u>some</u> neural network, provided our architecture is sufficiently general. However we must work with a compact space W of parameters, and hence networks of bounded depth, and with such a constraint there is no reason a priori to assume the true distribution is realisable in deep learning on real-world datasets.

For example, consider the training cross-entropy loss curves for GPT-3 from [OA2]:

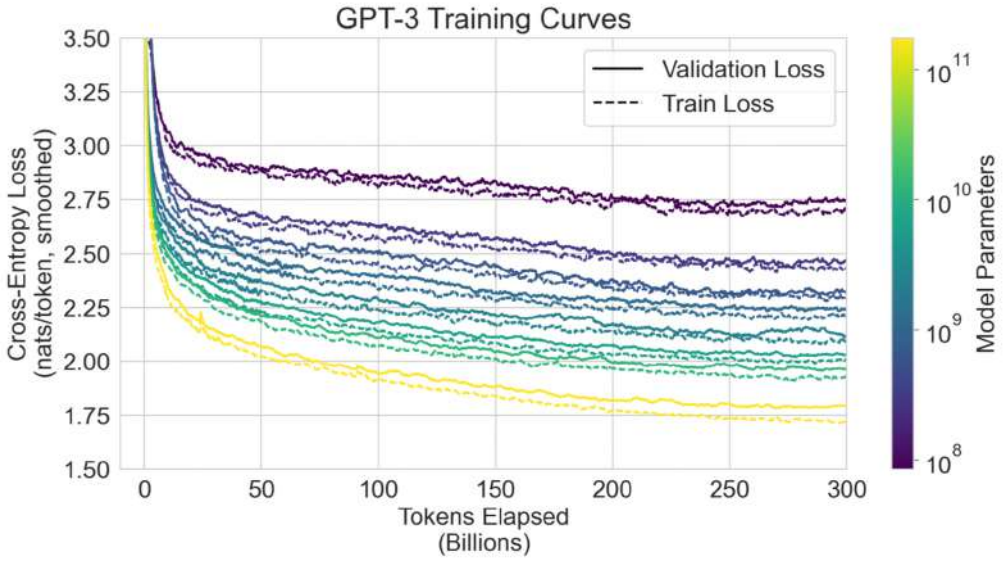


Figure 4.1: GPT-3 Training Curves We measure model performance during training on a deduplicated validation split of our training distribution. Though there is some gap between training and validation performance, the gap grows only minimally with model size and training time, suggesting that most of the gap comes from a difference in difficulty rather than overfitting.

References

[W] S. Watanabe "Algebraic geometry and statistical learning theory" 2009.

[WBIC] S. Watanabe "A widely applicable Bayesian information criterion" 2013

[WA] S. Watanabe "Asymptotic learning curve and renormalizable condition in singular learning theory" 2010

[GBC] I. Goodfellow, Y. Bengio, A. Courville "Deep learning".

[OA1] J. Kaplan et al "Scaling laws for neural language models" OpenAI 2020.

[OA2] T. Brown et al "Language models are few-shot learners" OpenAI 2020.

[H] J. Hestness et al "Deep learning scaling is predictable, empirically" Baidu 2017.

[HS] B. Hanin, M. Selke "Approximating continuous functions by ReLU nets of minimal width" 2017.

[L] F. Li "ImageNet: where have we gone? where are we going?" ACM talk learning.acm.org/techtalks/ImageNet.

[JB] J. Brownlee "A gentle introduction to the ImageNet challenge" search on Google.

[I] J. Deng et al "ImageNet: a large-scale hierarchical image database" CVPR 2009.