Proceedings of the International Conference on Statistics: Theory and Applications (ICSTA'19)
Lisbon, Portugal – August 13 - 14, 2019
Paper No. 34
DOI: 10.11159/icsta19.34

On Fitting Complex Models to Noisy Data

Mu Zhu
Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1
mu.zhu@uwaterloo.ca

Abstract - Deep learning has been remarkably successful at solving certain problems; for other problems, however, its success has been much more limited. A generic explanation for this phenomenon is as follows: if the model we use to fit the data is not complex enough for the underlying signal, we will never fully catch it, even without any noise in the data; if the model is more complex than necessary, we can recover the signal only when there is little to no noise in the data; with a noisy data set, our ability to recover the signal deteriorates as we use more and more complex models.

Keywords: bias-variance decomposition; deep learning; entanglement; functional estimation; integrated mean-squared error; interaction effects; orthonormal basis; penalized regression splines.

1. Introduction

Recent breakthroughs in machine learning—most notably, deep neural networks (DNNs) [1]—are challenging the discipline of statistics. A captivating landmark achievement is the defeat of human professional Go champion, Lee Sedol, by a computer program called AlphaGo, which makes heavy use of the DNN technology [2]. It has prompted many to rejoice, or lament, that "[d]eep learning is killing every problem in [artificial intelligence]" [3].

In a recent public lecture [4] given at the University of Toronto, Professor Frank Harrell argued that these DNN models are most fitting for problems with relatively low noise levels; for problems with relatively high noise levels, they haven't yet delivered the kind of spectacular success that everyone is expecting. According to Professor Harrell, learning to evaluate different moves in the game of Go is a "low noise" problem. The game is highly complex due to the exploding number of possible configurations but, given any particular configuration, the value of each legal move is more or less deterministic; there is little uncertainty about it.

The difference between {variation, complexity} and {uncertainty, noise} can be confusing indeed. They are, of course, not the same thing. For example, learning to recognize images of cats is another "low noise" problem of the kind that Professor Harrell had in mind. There is a lot of variation in this problem, because there are many different types of cats and even the same cat can look very different from different angles. In order to recognize such a great variety of cats, we need a relatively complex model. But there is a fundamental difference between

(a) having many clear images of different cats, and
(b) having many blurred images of the same cat.

In scenario (a), the signal itself is complex, containing a lot of variation, but upon reception it is mostly loud and clear. In scenario (b), the signal itself is simpler, but it is always badly corrupted when we receive it. Professor Harrell's point is essentially that complex models such as DNNs are really only suitable for scenario (a). In this short paper, I describe some simple analysis to support Professor Harrell's point. The results themselves are not really new; it's the way they are presented that may be helpful.

2. Experiment and Observations

In each panel of Fig. 1, a model with a certain level of complexity (as measured by the parameter, df) is fitted to data generated from the same underlying signal plus different levels of noise (as measured by the parameter, σ). From Fig. 1, we can make a few general observations:

(O1) If the model we use to fit the data is not complex enough for the underlying signal (two left columns), we will never fully capture it, even if there isn't any noise in the data at all (top row).

(O2) If the model is more complex than necessary (right column), we can still recover the signal when there is no noise in the data (top row).

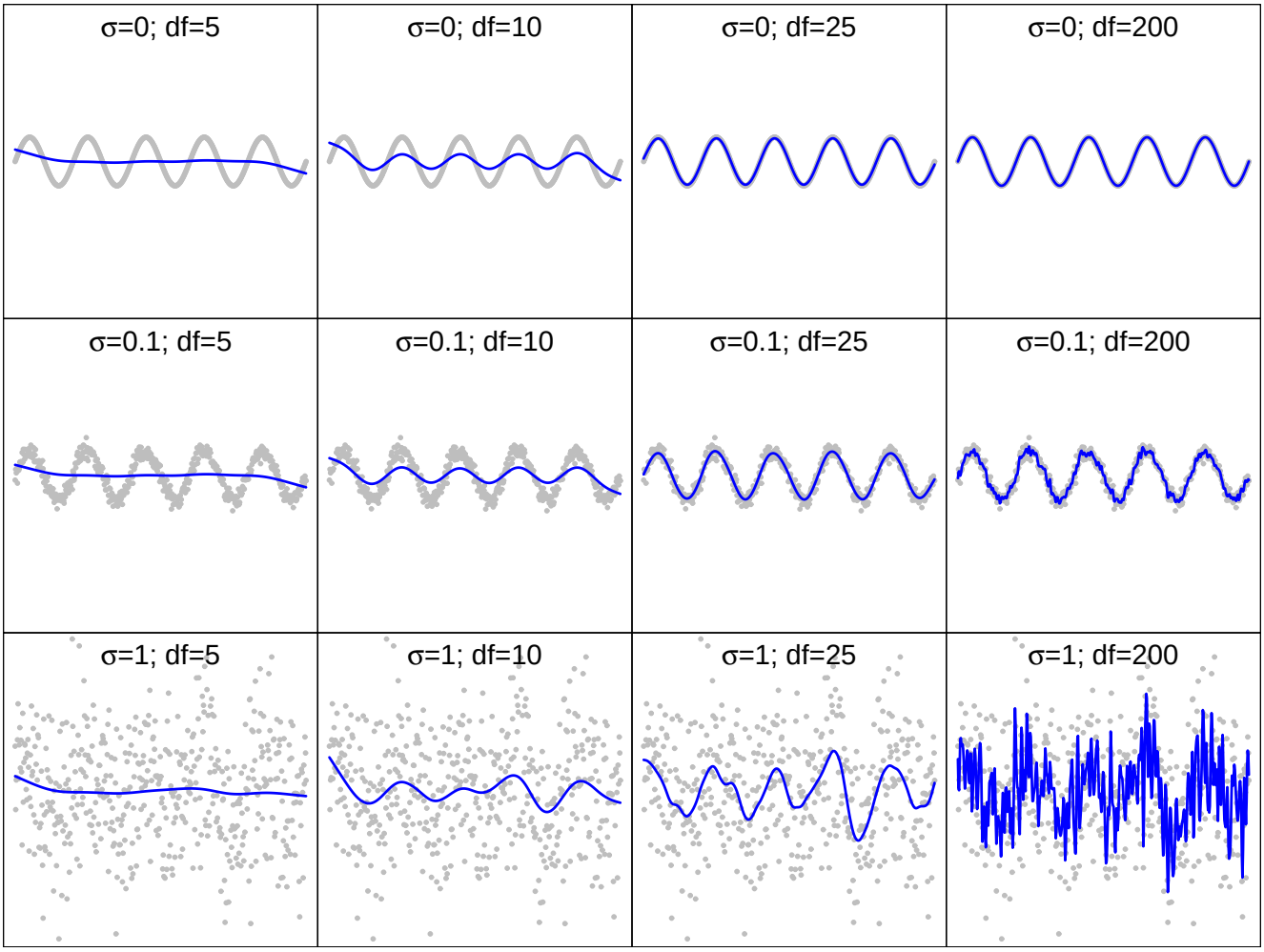

Fig. 1: Models with different levels of complexity (as measured by the parameter, df) are fitted to data generated from the same underlying signal plus different levels of noise (as measured by the parameter, σ). For $n = 500$ equally spaced points, $x_1, x_2, ..., x_n \in [0, 2\pi]$, responses are generated according to $y_i = f(x_i) + \varepsilon_i$, with ε_i being i.i.d. from $N(0, \sigma^2)$, and $f(x) = (0.5)\sin(5x)$. A penalized regression spline—see, e.g., [5, 6]—with a fixed set of 250 knots and a specific degree of freedom (df) is then fitted to the resulting data set, $\{(x_i, y_i)\}_{i=1}^n$.

(O3) But, as the data become noisier (moving down the rows while staying in the right column), we start to overfit.

(O4) Moreover, if the noise level is low (middle row), we do not suffer too badly; but if the noise level is high (bottom row), we can easily have a disaster.

Even though the observations (O1)-(O4) are hereby based upon a particular experiment, they are in fact not specific to such a toy example alone, and we can try to express them more formally.

3. Conjectures

For any function $f : \mathbb{R}^p \mapsto \mathbb{R}$, let $C(f)$ be a measure of its complexity. For $i = 1, 2, ..., n$, suppose $y_i = f(x_i) + \varepsilon_i$, where $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{V}\text{ar}(\varepsilon_i) = \sigma^2$, $\mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$, and the function f is unknown. Let

$$\widehat{f}_{n,q} \quad = \quad \underset{C(g) \leq q}{\arg\min} \quad \sum_{i=1}^{n} [y_i - g(x_i)]^2$$

Table 1: How each observation (O1), (O2), (O3) and (O4) is supported by various elements of Theorem 1.

	Observation	Theorem	
$q < C(f)$	(O1)	(C1)	$L > 0$ even if $\sigma^2 = 0$
$q \geq C(f)$	(O2)	(C2)(i)	$h_n(q,0) \equiv 0$
	(O3)	(C2)(ii)	$h_n(q,\sigma^2) > 0$ for $\sigma^2 > 0$
	(O4)	(C2)(iii)	$h_n(q,\sigma^2) = q\sigma^2/n$

be an estimate of f based on fitting a function with complexity no more than q to the data $\{(x_i,y_i)\}_{i=1}^n$. Based on (O1)-(O4), it is tempting to make the following conjectures about

$$\text{IMSE}(\widehat{f}_{n,q}) \equiv \int \mathbb{E}(|f(x) - \widehat{f}_{n,q}(x)|^2)dx,$$

the integrated mean-squared error of $\widehat{f}_{n,q}$:

(C1) If $q < C(f)$, then there exists an irreducible lower bound $L > 0$ such that $\text{IMSE}(\widehat{f}_{n,q}) \geq L$ even if $\sigma^2 = 0$ and/or $n \to \infty$.

(C2) If $q \geq C(f)$, then there exists a function $h_n(q,\sigma^2)$ such that $\text{IMSE}(\widehat{f}_{n,q}) \asymp h_n(q,\sigma^2)$, where

(i) $h_n(q,0) \equiv 0$,

(ii) $h_n(q,\sigma^2) > 0$ for $\sigma^2 > 0$, and

(iii) $h_n(q,\sigma^2)$ is non-decreasing in both q and σ^2.

The technical nuance here lies in the exact definition of the complexity measure $C(\cdot)$ and the precise form of the function $h_n(\cdot,\cdot)$.

4. Extra Assumptions and Theorem

While a general proof of (C1)-(C2) for any definition of $C(f)$ may be difficult to obtain, it is not too difficult to prove a specific version of them by introducing two additional assumptions:

(A1) There exists an orthonormal basis $\mathscr{B} = \{\varphi_1(x), \varphi_2(x), ...\}$ such that

$$f(x) = \sum_{j=1}^{q_*} \beta_j \varphi_j(x).$$

(A2) The set of basis functions, $\mathscr{B}$, is known (but not the number q_*).

Under (A1), we can define $C(f) = q_*$ to be the number of basis functions in $\mathscr{B}$ required to express f. For any fixed choice of q, (A2) allows us to obtain $\widehat{f}_{n,q}$ simply by regressing y_i onto $\varphi_1(x_i), ..., \varphi_q(x_i)$. These simplifications make it possible to turn (C1)-(C2) into a theorem.

Theorem 1. *Under (A1) and (A2), (C1) and (C2) hold with* $h_n(q,\sigma^2) = q\sigma^2/n$.

A proof of Theorem 1 is given in the Appendix. It is trivial to verify that $h_n(q,\sigma^2) = q\sigma^2/n$ satisfies the three requirements (i)-(iii) laid out in (C2). Table 1 explains how various elements of Theorem 1 can be seen to provide theoretical support for each of the observations (O1)-(O4) we have made earlier.

5. Discussion

Do the two extra assumptions (A1)-(A2) render the result practically irrelevant? I will argue that they do not. Assumption (A1) is not terribly unrealistic; it is relatively common in the theoretical literature. Assumption (A2) is clearly the main culprit. However, one can argue that, in reality, the estimation error cannot be smaller than if we already know the dictionary, $\mathscr{B}$. In other words, one could almost rely on (A1) alone and "elevate" (C1)-(C2) into a theorem by simply revising (C2) to say

$$\text{IMSE}(\widehat{f}_{n,q}) \geq \frac{q\sigma^2}{n}$$

for $q \geq C(f)$, instead of

$$\text{IMSE}(\widehat{f}_{n,q}) \asymp h_n(q,\sigma^2)$$

for some unspecific function $h_n(\cdot,\cdot)$ satisfying (i)-(iii).

Such a revision hardly affects our main qualitative conclusion. Instead of quantifying the estimation error *itself*, the revised statement merely quantifies a *lower bound* on the error, but it still lends the same theoretical support for observations (O2)-(O4). Moreover, the practical implications remain unchanged as well. That is, when there is little noise in the data, it pays to use a complex model [observation (O1) and (O2)]—this explains why deep learning has been so successful at solving certain problems; however, when data are very noisy, there are definitely good reasons why we should be cautious about, or even eschew, fitting very complex models [observation (O4)].

The expression, $h_n(q,\sigma^2) = q\sigma^2/n$, also shows that there is *entanglement* between complexity (q) and noise (σ^2). If we are willing to tolerate a slight abuse of statistical jargon, we can also say that there is an *interaction effect* between these two factors on estimation error.

6. Conclusion

I have provided some simple analysis to support Professor Harrell's point of view that the deep learning (DL) community seem to be betting on high complexity and low noise. Specifically, if the signal itself is complex in the sense that $C(f)$ is large, then using a simple, or shallow, model is not enough; if the noise level σ^2 is relatively low, then using a complex, or deep, model won't hurt too badly. For the kind of problems that the DL community tend to focus on, they definitely appear to be betting in the right direction. However, the success of their bets does not automatically imply that all estimation problems are cast in such "low noise" environment. For many other problems, we may very well be required to bet in the other direction.

Acknowledgments

The author's research is supported partially by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Discovery Grant No. RGPIN-2016-03876.

References

[1] Y. LeCun, Y. Bengio and G. Hinton, G, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.

[2] 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, pp. 484-489, 2016.

[3] E. Gibney, "Google AI algorithm masters ancient game of Go," *Nature*, vol. 529, pp. 445-446, 2016.

[4] F. E. Harrell, "Musings on statistical models versus machine learning in health research," Public Lecture, Dalla Lana School of Public Health, University of Toronto, May 2, 2019.

[5] P. Green and B. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, 1994.

[6] T. J. Hastie, R. J. Tibshirani and J. H. Friedman, *The Elements of Statistical Learning: Data-mining, Inference and Prediction*. Springer-Verlag, 2001.

Appendix

We prove (C1)-(C2) under the additional assumptions, (A1)-(A2). Throughout the proof, we will simply write $\widehat{f}$ instead of $\widehat{f}_{n,q}$. Given q, let $y = (y_1, y_2, \cdots, y_n)^\top$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^\top$,

$$\Phi = \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_q(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_q(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots & \varphi_q(x_n) \end{bmatrix}, \quad \text{and} \quad \varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \\ \varphi_q(x) \end{bmatrix}.$$

Under (A2), we simply regress y onto Φ, obtain

$$\widehat{\beta} = (\Phi^\top \Phi)^{-1} \Phi^\top y,$$

and estimate f by

$$\widehat{f}(x) = \sum_{j=1}^{q} \widehat{\beta}_j \varphi_j(x) = [\varphi(x)]^\top \widehat{\beta}.$$

The proof primarily consists of calculating the IMSE of $\widehat{f}$ using the well-known bias-variance decomposition:

$$\int \mathbb{E}(|f(x) - \widehat{f}(x)|^2) dx = \int \mathbb{B}\text{ias}^2[\widehat{f}(x)] dx + \int \mathbb{V}\text{ar}[\widehat{f}(x)] dx. \tag{1}$$

Step 0

That $\{\varphi_1, \varphi_2, ...\}$ is an orthonormal basis means

$$\frac{1}{n} \sum_{i=1}^{n} \varphi_j^2(x_i) \asymp \int \varphi_j^2(x) dx = 1 \quad \forall \quad j \tag{2}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \varphi_j(x_i) \varphi_k(x_i) \asymp \int \varphi_j(x) \varphi_k(x) dx = 0 \quad \forall \quad j \neq k, \tag{3}$$

for relatively large n. Hence, we have the approximation

$$\Phi^\top \Phi \asymp nI. \tag{4}$$

Step 1

First, we compute the variance part of (1). For fixed x,

$$\begin{aligned} \mathbb{V}\text{ar}[\widehat{f}(x)] &= [\varphi(x)]^\top \mathbb{V}\text{ar}(\widehat{\beta}) [\varphi(x)] \\ &= [\varphi(x)]^\top \left[\sigma^2 (\Phi^\top \Phi)^{-1} \right] [\varphi(x)] \\ &\asymp \frac{\sigma^2}{n} \sum_{j=1}^{q} \varphi_j^2(x), \end{aligned}$$

using the approximation (4). Hence,

$$\int \mathbb{V}\text{ar}[\widehat{f}(x)] dx \asymp \frac{\sigma^2}{n} \sum_{j=1}^{q} \int \varphi_j^2(x) dx = \frac{q\sigma^2}{n}. \tag{5}$$

Step 2

Next, we compute the bias part of (1). First, suppose $q < q_* \equiv C(f)$. Define a few "obvious" quantities—$\{\Phi_m, \varphi_m\}$, where the subscript "m" stands for "missing"; $\{\Phi_*, \varphi_*\}$; and $\{\beta, \beta_m, \beta_*\}$—in such a way that

$$\Phi_* = [\ \Phi \quad \Phi_m\ ], \quad \varphi_*(x) = \left[\begin{array}{c} \varphi(x) \\ \varphi_m(x) \end{array}\right], \quad \text{and} \quad \beta_* = \left[\begin{array}{c} \beta \\ \beta_m \end{array}\right].$$

Then,

$$y = \Phi_*\beta_* + \varepsilon = [\ \Phi \quad \Phi_m\ ] \left[\begin{array}{c} \beta \\ \beta_m \end{array}\right] + \varepsilon$$

and, for fixed x, we can obtain

$$\begin{aligned}
\mathbb{B}\text{ias}[\widehat{f}(x)] &= \mathbb{E}[\widehat{f}(x)] - f(x) \\
&= [\varphi(x)]^\top \mathbb{E}(\widehat{\beta}) - [\varphi_*(x)]^\top \beta_* \\
&= [\varphi(x)]^\top (\Phi^\top \Phi)^{-1} \Phi^\top \mathbb{E}(y) - [\varphi_*(x)]^\top \beta_* \\
&= [\varphi(x)]^\top (\Phi^\top \Phi)^{-1} \Phi^\top (\Phi\beta + \Phi_m\beta_m) - \left\{ [\varphi(x)]^\top \beta + [\varphi_m(x)]^\top \beta_m \right\} \\
&= [\varphi(x)]^\top (\Phi^\top \Phi)^{-1} \Phi^\top \Phi_m \beta_m - [\varphi_m(x)]^\top \beta_m.
\end{aligned}$$

However, by (2)-(3), we have $\Phi^\top \Phi_m \asymp 0$, so

$$\mathbb{B}\text{ias}[\widehat{f}(x)] \asymp -[\varphi_m(x)]^\top \beta_m.$$

But the fact that $C(f) = q_* > q$ means not all elements of β_m can be zero. Thus, we have found a quantity,

$$L = \int \mathbb{B}\text{ias}^2[\widehat{f}(x)]dx \asymp \int \left\{ [\varphi_m(x)]^\top \beta_m \right\}^2 dx > 0, \tag{6}$$

which is strictly larger than zero. Combining (5) and (6), we see that, when $q < C(f)$,

$$\text{IMSE}(\widehat{f}) \asymp L + \frac{q\sigma^2}{n} > 0,$$

even if $\sigma^2 = 0$ and/or $n \to \infty$.

Step 3

Now, suppose $q \geq q_* \equiv C(f)$. To proceed, re-define the quantities $\{\Phi_m, \varphi_m\}$, $\{\Phi_*, \varphi_*\}$ and $\{\beta, \beta_m, \beta_*\}$ so that

$$\Phi = [\ \Phi_* \quad \Phi_m\ ], \quad \varphi(x) = \left[\begin{array}{c} \varphi_*(x) \\ \varphi_m(x) \end{array}\right], \quad \text{and} \quad \beta = \left[\begin{array}{c} \beta_* \\ 0 \end{array}\right].$$

Then,

$$y = \Phi_*\beta_* + \varepsilon = [\ \Phi_* \quad \Phi_m\ ] \left[\begin{array}{c} \beta_* \\ 0 \end{array}\right] + \varepsilon = \Phi\beta + \varepsilon$$

and, for fixed x, we can obtain

$$\begin{aligned}
\mathbb{B}\text{ias}[\widehat{f}(x)] &= \mathbb{E}[\widehat{f}(x)] - f(x) \\
&= [\varphi(x)]^\top \mathbb{E}(\widehat{\beta}) - [\varphi_*(x)]^\top \beta_* \\
&= [\varphi(x)]^\top (\Phi^\top \Phi)^{-1} \Phi^\top \mathbb{E}(y) - [\varphi_*(x)]^\top \beta_* \\
&= [\varphi(x)]^\top (\Phi^\top \Phi)^{-1} \Phi^\top \Phi\beta - [\varphi_*(x)]^\top \beta_* \\
&= [\varphi(x)]^\top \beta - [\varphi_*(x)]^\top \beta_* \\
&= \left\{ [\varphi_*(x)]^\top \beta_* + [\varphi_m(x)]^\top 0 \right\} - [\varphi_*(x)]^\top \beta_* \\
&= 0,
\end{aligned}$$

which means

$$\int \mathbb{B}\text{ias}^2[\widehat{f}(x)]dx = 0 \tag{7}$$

as well. Combining (5) and (7), we see that, when $q \geq C(f)$,

$$\text{IMSE}(\widehat{f}) \asymp \frac{q\sigma^2}{n},$$

which is (i) equal to zero when $\sigma^2 = 0$, (ii) strictly positive if $\sigma^2 > 0$, and (iii) non-decreasing in both (q, σ^2).