How badly do Interpolating Polynomials overfit?

Simon Segert

May 5, 2024

1 Introduction

It is often claimed in intro stats/machine learning treatments that polynomial interpolation leads to overfitting. This is not strictly true, but a more specific claim is: namely, fitting a polynomial to exactly pass through each data point does lead to overfitting (this is called the Lagrange interpolant). But under what conditions is this true, and how can we quantify how bad the overfitting really is?

We provide here a rigorous analysis showing that **the variance of the Lagrange polynomial fit for evenly-spaced points grows exponentially as a function of** n. We also provide a follow-up analysis to show that if the x points are sampled from some distribution μ supported on a finite interval, then **the variance still grows exponentially, unless** μ **is the arcsine distribution**.

Thus we see that the overfitting is actually of an extreme kind: not only does the variance of the interpolants not go to zero, it actually grows very quickly as more data points are added. We also get an unexpected insight that the arcsine distribution is the "friendliest" distribution for the Lagrange interpolants.

2 Analysis of Evenly-spaced points

As a first step, we consider the following setting: there is some function $f : [0, 1] \to \mathbb{R}$ and we have access to n noisy observations at equally-spaced points in the unit interval. That is, we know $(x_i, f(x_i)+\epsilon_i)$ where $x_i = i/n$, $n = 1, \ldots, n$ and $\epsilon_i \sim N(0, 1)$ is iid normal noise. Let $\hat{f}_\epsilon : [0, 1] \to \mathbb{R}$ be some estimator that depends on these observations. The notation indicates that we regard x_i as fixed and ϵ_i as random. Thus we can also think of $\hat{f}$ as a random function, with the randomness coming from the distribution of ϵ. We define the **Variance** of the estimator as

$$V(\hat{f}) = \int_0^1 Var_\epsilon f_\epsilon(x)dx$$

that is, the integrated variance of the values of $\hat{f}$ at each point in the interval. By the standard Bias-variance decomposition, this provides a lower bound on the mean-squared error of the estimator when evaluated on a new point $x \in [0, 1]$.

We will be concerned with $\hat{f}$ taking the form of the Lagrange interpolant:

Definition 2.0.1. *If $(x_i, y_i), i = 1, \ldots, n$ are pairs of points with $x_i, y_i \in \mathbb{R}$, then the **Lagrange interpolant** is the unique polynomial p of degree $n-1$ such that $p(x_i) = y_i, \forall i$.*

We can now state the first theorem.

Theorem 2.1. *Let $\hat{f}$ denote the Lagrange interpolant of the points $(x_i, f(x_i) + \epsilon_i)$ where $x_i = i/n$, $i = 1, 2, \ldots, n$. Then $V(\hat{f}) \geq Ce^{n\delta}$ where C is some positive constant that does not depend on n or f, and $\delta = 3/2 - \log 4 > 0$.*

Proof. Introduce the notation $pow(x) = (1, x, \ldots, x^{n-1})$. Then the interpolant $\hat{f}_\epsilon$ can be written in the form

$$\hat{f}_\epsilon(x) = \langle w, pow(x) \rangle$$

for some $w \in \mathbb{R}^n$. Because the Lagrange interpolant passes through each datapoint, we must have $f(x_i) + \epsilon_i = \hat{f}_\epsilon(x_i) = (Vw)_i$ where the Vandermonde matrix V is defined by $V_{ij} = x_i^{j-1}$, $i, j = 1, \ldots, n$ So we have the matrix formula $w = V^{-1}y = V^{-1}f(x) + V^{-1}\epsilon$ which gives

$$\hat{f}_\epsilon(x) = \langle V^{-1}f(x), pow(x) \rangle + \langle V^{-1}\epsilon, pow(x) \rangle$$

Since the first term does not depend on ϵ,

$$\begin{aligned} Var_\epsilon \hat{f}_\epsilon(x) &= Var_\epsilon \langle V^{-1}\epsilon, pow(x) \rangle \\ &= Var_\epsilon \langle \epsilon, (V^{-1})^T pow(x) \rangle \\ &= pow(x)^T V^{-1} (V^{-1})^T pow(x) \\ &= Tr((V^T V)^{-1} pow(x) pow(x)^T) \end{aligned}$$

Therefore

$$\begin{aligned} Var(\hat{f}) &= \int_0^1 Var_\epsilon \hat{f}_\epsilon(x) \\ &= \int_0^1 Tr((V^T V)^{-1} pow(x) pow(x)^T) dx \\ &= Tr((V^T V)^{-1} \int_0^1 pow(x) pow(x)^T dx) \\ &= Tr((V^T V)^{-1} H_n) \\ &= Tr(V^{-1} H_n (V^{-1})^T) \end{aligned}$$

where H_n is the Hilbert matrix $(H_n)_{ij} = 1/(i + j - 1)$, $i, j = 1, \ldots, n$.

Thus, we have expressed the variance as the trace of a positive definite matrix (the Hilbert matrix is positive definite, and positive definite matrices remain so under conjugation). In general, any positive definite matrix M satisfies

$Tr(M) \geq n \det(M)^{1/n} \geq det(M)^{1/n}$, as can be seen by applying the Arithmetic-Geometric mean inequality to the eigenvalues of M.

In our case this implies

$$\begin{aligned} Var(\hat{f}) &\geq det(H_n)^{1/n}|det(V)|^{-2/n} \qquad (1) \\ n \log Var(\hat{f}) &\geq \log\det(H_n) - 2\log|\det(V)| \qquad (2) \end{aligned}$$

We consider these determinants one by one. The Hilbert determinant has a well-known [1] asymptotic expansion:

$$\log\det(H_n) \sim -n^2\log(4) + O(n) \qquad (3)$$

For the Vamdermonde determinent, we have the well-known formula $det(V) = \prod_{i<j} x_j - x_i$ which in our case reduces to

$$\begin{aligned} det(V) &= \prod_{i<j}(j-i)/n \\ &= n^{-(n^2-n)/2}\prod_{i<j} j-i \\ &= n^{-(n^2-n)/2}\prod_{i=1}^{n-1} i! \end{aligned}$$

Similarly to the relation between factorials and the Gamma function, the products of factorials are interpolated by a certain analytic function called the Barnes G function [2], G. More precisely, we have

$$G(n+1) = \prod_{i=1}^{n-1} i! \qquad (4)$$

when n is a positive integer. Furthermore, there is an analogue of Stirling's formula:

$$\log G(n+1) = n^2\log n/2 - 3n^2/4 + O(n) \qquad (5)$$

So the determinant goes like

$$\begin{aligned} \log\det|V| &= -\frac{n^2-n}{2}\log n + \frac{n^2}{2}\log n - \frac{3n^2}{4} + O(n) \qquad (6) \\ &= -\frac{3n^2}{4} + O(n) \qquad (7) \end{aligned}$$

Putting together Equations 2,3,7 we get

$$\begin{aligned} n\log Var(\hat{f}) &\geq \left(\frac{3}{2} - \log 4\right) n^2 + O(n) \\ \log Var(\hat{f}) &\geq \left(\frac{3}{2} - \log 4\right) n + O(1) \end{aligned}$$

□

[1] https://en.wikipedia.org/wiki/Hilbert_matrix

[2] https://en.wikipedia.org/wiki/Barnes_G-function

3 other distributions

In the previous section, we considered observation points that are exactly evenly spaced. However, a more realistic setup would be to have x_i be independent samples from some fixed distribution μ.

In this case, the interpolant $\hat{f}$ depends on both $\epsilon = \{\epsilon_i\}_i$ and $x = \{x_i\}_i$, and to the definition of the variance must accordingly be modified to marginalize over both sources of randomness:

$$Var(\hat{f}) := \mathbb{E}_{x_i \sim \mu} \int_0^1 Var_\epsilon \hat{f}_{x,\epsilon}(x) dx \tag{8}$$

Theorem 3.1. *Let μ be a distribution supported on the interval $[0,1]$ and let $\tilde{\mu}$ be the affinely-rescaled distribution that is supported on $[-1,1]$. Assume that $\tilde{\mu}$ is such that*

- *$\tilde{\mu}$ is absolutely continuous*
- $\int_{-1}^1 \left(\frac{d\tilde{\mu}}{dx}\right)^2 \sqrt{1-x^2} dx < \infty$
- *$\tilde{\mu}$ is not equal to the arcsine distribution $\frac{d\mu_{arcsin}}{dx} = \frac{1}{\pi\sqrt{1-x^2}}$*

If there are n observation points x_i which are sampled iid from μ, then the variance $Var(\hat{f})$ grows at least exponentially as a function of n. In fact, it grows at least as $O(e^{\delta n})$ where $\delta = -\int_0^1 \int_0^1 \log|x-y| d\mu(x) d\mu(y) - \log 4 > 0$.

We first prove the following key technical lemma:

Lemma 3.1. *Let $\tilde{\mu}$ be some distribution on $[-1,1]$ which satisfies the conditions of Theorem 3.1. Then*

$$-\int_{-1}^1 \int_{-1}^1 \log|x-y| d\tilde{\mu}(x) d\tilde{\mu}(y) > \log 2 \tag{9}$$

Proof. Recall that in general any function $f : [-1,1] \to \mathbb{R}$ such that $\int_{-1}^1 \frac{f(x)^2}{\sqrt{1-x^2}} dx < \infty$ can be expanded as a sum $\sum_{n \geq 0} c_n T_n(x)$ of Chebyshev polynomials. By the assumptions, this is satisfied by the function $\frac{d\tilde{\mu}}{dx}(x)\sqrt{1-x^2}$, therefore

$$\frac{d\tilde{\mu}}{dx}(x) = \sum_{n \geq 0} \frac{c_n T_n(x)}{\sqrt{1-x^2}} \tag{10}$$

where the series converges in the distributional sense (cf. [EK99] Ch. 7).

Define the functional

$$\tilde{J}(\tilde{\mu}) = -\int_{-1}^1 \int_{-1}^1 \log|x-y| d\tilde{\mu}(x) d\tilde{\mu}(y) \tag{11}$$

To find a minimum, we add a Lagrange multiplier to enforce the constraint $\int_{-1}^{1} d\tilde{\mu}(x) = 1$ and obtain the condition for a critical point:

$$\int_{-1}^{1} \log|x-y| d\tilde{\mu}(y) = \lambda \tag{12}$$

where λ is some constant.

We have the series expansion ([EK99]):

$$-\log|x-y| = \log 2 + \sum_{n\geq 1} \frac{2}{n} T_n(x) T_n(y) \tag{13}$$

By the orthogonality properties of Chebyshev polynomials, this implies that any solution of Equation 12 must be a multiple of the arcsine distribution; when combined with the normalization condition, it follows that the arcsine distribution is in fact the only critical point of $\tilde{J}$ within the class of probability distributions satisfying the hypotheses of the lemma.

The expansion in Equation 13 moreover implies that the kernel is strictly positive definite for any $\tilde{\mu}$ that admits an expansion as in Equation 10. Thus the minimum is strict.

The last thing to show is that $\tilde{J}(\mu_{arcsin}) = \log 2$. This follows immediately from Equation 13 and the orthogonality properties of Chebyshev polynomials. □

Proof. (of Theorem 3.1)

The argument given in the proof of Theorem 2.1 goes through exactly as before to obtain

$$n \log \int_0^1 Var_\epsilon \hat{f}_{x,\epsilon}(x) dx \geq \log\det H_n - 2\log|\det V(x)| \tag{14}$$

where the Vandermonde determinant is now with respect to the sampled points x_i.

The difference now is we also have to marginalize over x_i. Since log is a concave function, Jensen's inequality gives

$$\begin{aligned} \log V(\hat{f}) &= \log \mathbb{E}_x \int_0^1 Var_\epsilon \hat{f}_{x,\epsilon}(x) dx && (15) \\ &\geq \mathbb{E}_x \log \int_0^1 Var_\epsilon \hat{f}_{x,\epsilon}(x) dx && (16) \end{aligned}$$

Therefore

$$n \log V(\hat{f}) \geq \log\det H_n - 2\mathbb{E}_x \log|\det V(x)| \tag{17}$$

The log Vandermonde determinant is a sum of independent terms $\log|x_i - x_j|$ and is therefore in expectation given by

$$2\mathbb{E}_{x_i \sim \mu} \log|\det V(x)| = (n^2 - n) \int_0^1 \int_0^1 \log|x-y| d\mu(x) d\mu(y)$$

Summing up, we have the lower bound

$$n \log V(\hat{f}) \geq \log \det H_n + (n^2 - n) J(\mu)$$

where we have introduced the functional

$$J(\mu) := -\int_0^1 \int_0^1 \log |x - y| d\mu(x) d\mu(y)$$

To apply the lemma, it will be convenient to rescale the integral so the limits are instead $[-1, 1]$. By an affine change of variables we get

$$\begin{aligned} J(\mu) &= \log 2 - \int_{-1}^1 \int_{-1}^1 \log |x - y| d\tilde{\mu}(x) d\tilde{\mu}(y) && (18) \\ &:= \log 2 + \tilde{J}(\tilde{\mu}) && (19) \end{aligned}$$

where $\tilde{\mu}$ denotes the shifted and rescaled version of μ which is supported on $[-1, 1]$ rather than $[0, 1]$, and $\tilde{J}$ is the functional from the statement and proof of the Lemma.

In particular, since $\tilde{\mu}$ is not the arcsine distribution, we have $\tilde{J}(\tilde{\mu}) > \log 2$ by the Lemma, and in particular

$$J(\mu) > \log 2 + \log 2 = \log 4 \qquad (20)$$

Define $\delta_\mu = J(\mu) - \log 4 > 0$. Then,

$$\begin{aligned} n \log Var(\hat{f}) &\geq \log \det H_n - 2\mathbb{E}_{x_i \sim \mu} \log |\det V(x)| \\ &= \log \det H_n + (n^2 - n) J(\mu) \\ &= \log \det H_n + (n^2 - n)(\log 4 + \delta_\mu) \\ &= -n^2 \log 4 + n^2 (\log 4 + \delta_\mu) + O(n) \\ &= \delta n^2 + O(n) \end{aligned}$$

□

To be clear, this doesn't imply that the arcsine *doesn't* have exponentially growing variace, just that it is the only distribution not handled by this argument.

We can also get a few interesting insights from this argument:

- The functional $J(\mu)$ controls the growth rate of the variance
- The arcsine distribution is in some sense the "best case scenario" for the Lagrange interpolants

References

[EK99] R Estrada and R Kanwal. *Singular Integral Equations*. 1999.