The Weierstrass density theorem

Harald Hanche-Olsen
hanche@math.ntnu.no

Bernshteĭn[1] polynomials are ordinary polynomials written on the particular form

$$b(t) = \sum_{k=0}^{n} \beta_k \binom{n}{k} t^k (1-t)^{n-k}, \tag{1}$$

where $\beta_0, \dots, \beta_n$ are given coefficients.[2] The special case where each $\beta_k = 1$ deserves mention: Then the binomial theorem yields

$$\sum_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k} = \bigl(t + (1-t)\bigr)^n = 1. \tag{2}$$

We can show by induction on n that if $b = 0$ then all the coefficients β_n are zero. The base case $n = 0$ is obvious. When $n > 0$, a little bit of binomial coefficient gymnastics shows that the derivative of a Bernshteĭn polynomial can be written as another Bernshteĭn polynomial:

$$b'(t) = n \sum_{k=0}^{n-1} (\beta_{k+1} - \beta_k) \binom{n-1}{k} t^k (1-t)^{n-1-k}.$$

In particular, if $b = 0$ it follows by the induction hypothesis that all β_k are equal, and then they are all zero, by (2).

In other words, the polynomials $t^k(1-t)^{n-k}$, where $k = 0, \dots, n$, are linearly independent, and hence they span the $n+1$-dimensional space of polynomials of degree ≤ 1. Thus *all polynomials can be written as Bernshteĭn polynomials*, so there is nothing special about these – only about the way we write them.

To understand why Bernshteĭn polynomials are so useful, consider the individual polynomials

$$b_{k,n}(t) = \binom{n}{k} t^k (1-t)^{n-k}, \qquad k = 0, \dots, n. \tag{3}$$

If we fix n and t, we see that $b_{k,n}(t)$ is the probability of k heads in n tosses of a biased coin, where the probability of a head is t. The expected number of heads in such an experiment is nt, and indeed when n is large, the outcome is very likely to be near that value. In other words, most of the contributions to the sum in (1) come from k near nt. Rather than using statistical reasoning, however, we shall proceed by direct calculation – but the probability argument is still a useful guide.

[1] Named after Sergeĭ Natanovich Bernshteĭn (1880–1968). The name is often spelled “Bernstein”.

[2] When $n = 3$, we get a *cubic spline*. In this case, β_0, β_1, β_2 and β_3 are called the *control points* of the spline. In applications, they are usually 2- or 3-dimensional vectors.

1 Theorem. (Weierstrass) *The polynomials are dense in* $C[0,1]$.

This will follow immediately from the following lemma.

2 Lemma. *Let* $f \in C[0,1]$. *Let* b_n *be the Bernshteĭn polynomial*

$$b_n(t) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} t^k (1-t)^{n-k}.$$

Then $\|f - b_n\|_\infty \to 0$ *as* $n \to \infty$.

Proof: Let $t \in [0,1]$. With the help of (2) we can write

$$f(t) - b_n(t) = \sum_{k=0}^{n} \Big(f(t) - f\Big(\frac{k}{n}\Big)\Big)\binom{n}{k} t^k (1-t)^{n-k},$$

so that

$$|f(t) - b_n(t)| \le \sum_{k=0}^{n} \Big|f(t) - f\Big(\frac{k}{n}\Big)\Big|\binom{n}{k} t^k (1-t)^{n-k}, \tag{4}$$

We now use the fact that f is *uniformly continuous*: Let $\varepsilon > 0$ be given. There is then a $\delta > 0$ so that $|f(t) - f(s)| < \varepsilon$ whenever $|t - s| < \delta$. We now split the above sum into two parts, first noting that

$$\sum_{|k-nt|<n\delta} \Big|f(t) - f\Big(\frac{k}{n}\Big)\Big|\binom{n}{k} t^k (1-t)^{n-k} \le \varepsilon \tag{5}$$

(where we used $|f(t) - f(k/n)| < \varepsilon$, and then expanded the sum to all indexes from 0 to n and used (2)). To estimate the remainder, let $M = \|f\|_\infty$, so that

$$\sum_{|k-nt|\ge n\delta} \Big|f(t) - f\Big(\frac{k}{n}\Big)\Big|\binom{n}{k} t^k (1-t)^{n-k} \le 2M \sum_{|k-nt|\ge n\delta} \binom{n}{k} t^k (1-t)^{n-k}. \tag{6}$$

To finish the proof, we need to borrow from the Chebyshev inequality in order to show that the latter sum can be made small. First we find

$$\sum_{k=0}^{n} k\binom{n}{k} t^k (1-t)^{n-k} = nt \sum_{k=0}^{n-1} \binom{n-1}{k} t^k (1-t)^{n-1-k} = nt. \tag{7}$$

(Rewrite the binomial coefficient using factorials, perform the obvious cancellation using $k/k! = 1/(k-1)!$, put nt outside the sum, change the summation index, and use (2).) Next, using similar methods,

$$\sum_{k=0}^{n} k(k-1)\binom{n}{k} t^k (1-t)^{n-k} = n(n-1)t^2 \sum_{k=0}^{n-2} \binom{n-2}{k} t^k (1-t)^{n-2-k} = n(n-1)t^2.$$

Adding these two together, we get

$$\sum_{k=0}^{n} k^2 \binom{n}{k} t^k (1-t)^{n-k} = nt\big((n-1)t+1\big). \tag{8}$$

Finally, using (2), (7) and (8), we find

$$\sum_{k=0}^{n} (nt-k)^2 \binom{n}{k} t^k (1-t)^{n-k} = (nt)^2 - 2(nt)^2 + nt\big((n-1)t+1\big) = nt(1-t).$$

The most important feature here is that the n^2 terms cancel out. We now have

$$\begin{aligned} nt(1-t) &\geq \sum_{|k-nt|\geq n\delta} (nt-k)^2 \binom{n}{k} t^k (1-t)^{n-k} \\ &\geq (n\delta)^2 \sum_{|k-nt|\geq n\delta} \binom{n}{k} t^k (1-t)^{n-k}, \end{aligned}$$

so that

$$\sum_{|k-nt|\geq n\delta} \binom{n}{k} t^k (1-t)^{n-k} \leq \frac{t(1-t)}{n\delta^2} \leq \frac{1}{4n\delta^2}. \tag{9}$$

Combining (4), (5), (6) and (9), we end up with

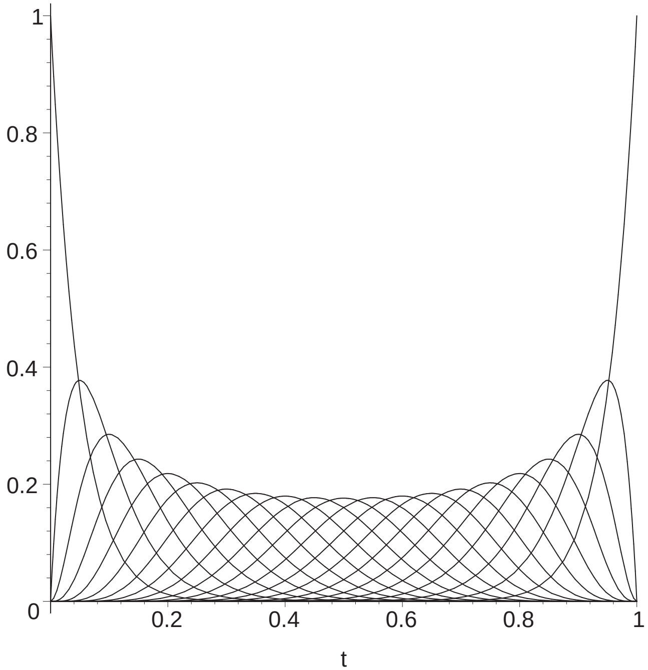
$$|f(t) - b_n(t)| < \varepsilon + \frac{M}{2n\delta^2}, \tag{10}$$

which can be made less than 2ε by choosing n large enough. More importantly, this estimate is independent of $t \in [0,1]$. ∎

One final remark: There is of course nothing magical about the interval $[0,1]$. Any closed and bounded interval will do. If $f \in C[a,b]$ then $t \mapsto f\big((1-t)a+tb\big)$ belongs to $C[0,1]$, and this operation maps polynomials to polynomials and preserves the norm. So the Weierstrass theorem works equally well on $C[a,b]$.

The Stone–Weierstrass theorem is a bit more difficult: It replaces $[0,1]$ by any compact set X and the polynomials by any algebra of functions which separates points in X and has no common zero in X. (This theorem assumes real functions. If you work with complex functions, the algebra must also be closed under conjugation. But the complex version of the theorem is not much more than an obvious translation of the the real version into the complex domain.) One proof of the general Stone–Weierstrass theorem builds on the Weierstrass theorem. More precisely, the proof needs an approximation of the absolute value $|t|$ by polynomials in t, uniformly for t in a bounded interval.

An amusing (?) diversion. Any old textbook on elementary statistics shows pictures of the binomial distribution, i.e., $b_{k,n}(t)$ for a given n and t; see (3). But it can be interesting to look at this from a different angle, and consider each term as a function of t. Here is a picture of all these polynomials, for $n = 20$:



We may note that $b_{k,n}(t)$ has its maximum at $t = k/n$, and $\int_0^1 b_{k,n}(t)\,dt = 1/(n+1)$. In fact, $(n+1)b_{k,n}$ is the probability density of a beta-distributed random variable with parameters $(k+1, n-k+1)$. Such variables have standard deviation varying between approximately $1/(2\sqrt{n})$ (near the center, i.e., for $k \approx n/2$) and $1/n$ (near the edges). Compare this with the distance $1/n$ between the sample points.

It is tempting to conclude that polynomials of degree n can only do a good job of approximating a function which varies on a length scale of $1/\sqrt{n}$.

We can see this, for example, if we wish to estimate a Lipschitz continuous function f, say with $|f(t) - f(s)| \le L|t - s|$. Put $\varepsilon = L\delta$ in (10) and then determine the δ that gives the best estimate in (10), to arrive at $|f(t) - b_n(t)| < \frac{3}{2}M^{1/3}(L^2/n)^{2/3}$. So the n required for a given accuracy is proportional to L^2, in accordance with the analysis in the previous two paragraphs.

Reference: S. N. Bernshteĭn: A demonstration of the Weierstrass theorem based on the theory of probability. *The Mathematical Scientist* **29**, 127–128 (2004).

By an amazing coincidence, this translation of Bernshteĭn's original paper from 1912 appeared recently. I discovered it after writing the current note.