The Stone-Weierstrass Theorem

Bill Casselman
University of British Columbia
cass@math.ubc.ca

A basic theme in representation theory is to approximate various functions on a space by simpler ones. Some well known examples are the approximation of functions on the interval $[0,1]$ by polynomials or that of approximation of functions on the unit circle by trigonometrical polynomials. A fundamental tool in all these results is a very general result about the approximation of continuous functions on compact Hausdorff spaces. In proving the general result it is useful to show first what happens for the simplest case of functions on $[0,1]$, which is as well known due to Weierstrass. I'll give below a satisfactorily explicit version of Weierstrass' theorem due to the Russian mathematician Sergei Bernstein, and in fact by following his own argument fairly closely. But in order to appreciate Bernstein's construction, it is useful to understand first that the naive approach to the problem of approximating continuous functions on $[0,1]$ doesn't work, and that's what I'll begin with.

1. Lagrange's interpolation

Given a continuous function f on $[0,1]$, a formula due to Lagrange produces a polynomial $L_{f,n}(x)$ that interpolates f exactly at the points m/n for $m = 0$ to n. First, for each m in that range define

$$L_{m,n}(x) = \prod_{0 \le k \le n, k \ne m} \frac{x - k/n}{m/n - k/n}\,.$$

This satisfies

$$L_{m,n}(j/n) = \begin{cases} 1 & j = m \\ 0 & \text{otherwise} \end{cases}$$

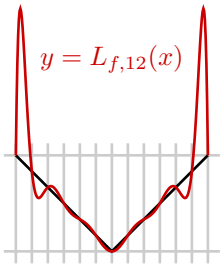
and hence if we choose $x_j = j/n$

$$L_{f,n}(x) = \sum_0^n f(k/n) L_{k,n}(x)$$

satisfies

$$L_{f,n}(k/n) = f(k/n) \quad (0 \le k \le n)\,.$$

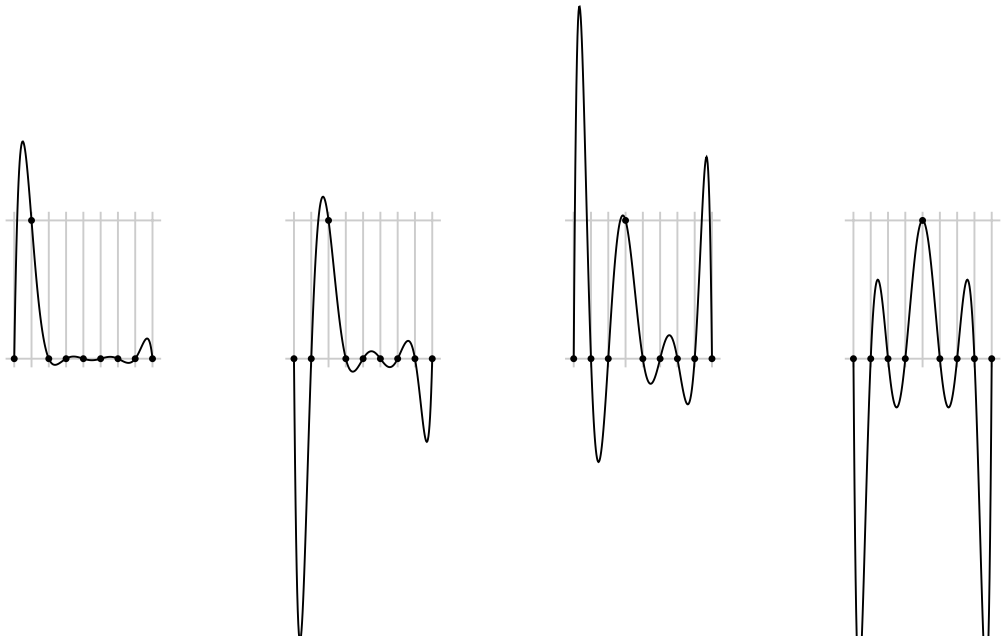
Lagrange's interpolating polynomial $L_{f,n}(x)$ therefore agrees exactly with $f(cx)$ at all points k/n. This seems like a good start, but in fact it does a poor job of approximating arbitrary continuous functions away from the interpolation points. Here, for example, is the graph of $L_{f,16}(x)$ for the function $f(x) = |x - 0.5|$ in the interval $[0,1]$:



$f(x) = |x - 1/2|$

More examples would show that the behaviour of P_n at the end points only becomes wilder as n increases.

You can see what the problem is, at least roughly—Lagrange's polynomial doesn't behave *locally*—i.e. singular behaviour of the function at a point can affect behaviour of the approximation far away from it. In these examples, the break at the middle of the interval ($x = 1/2$) causes severe oscillation at the endpoints ($x = 0, 1$) and as the number of interpolation points increases so does this oscillation. The following figures, which graph $L_{m,8}(x)$ for $m = 1$ to 4, exhibit the difficulty clearly.



The Lagrange polynomials are useless for demonstrating Weierstrass' theorem, and in fact what they demonstrate is that the theorem is subtle.

One way around the difficulty is to use Lagrange's formula at points that are not evenly spaced in $[0, 1]$. The optimal choice leads to approximation by Chebyshev polynomials, which often works well. But in the next section we'll see a simple, intuitive method with much appeal.

2. Bernstein polynomials

If $\beta = (\beta_k)$ for $k = 0$ to n is any sequence of real numbers, the associated **Bernstein polynomial** is

$$B_\beta(x) = \sum_{k=0}^{n} \beta_k \binom{n}{k} x^k (1-x)^{n-k} .$$

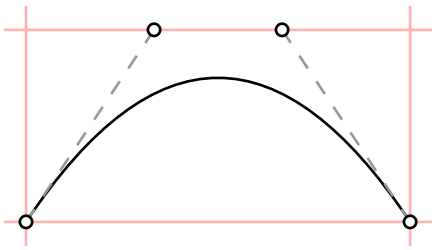
If all the β_k are equal to a single constant β, for example, we get the constant function

$$B_n(x) = \beta\big((1-t)+t\big)^n = \beta .$$

In other low degrees:

$$\begin{aligned} B_1(x) &= \beta_0(1-x) + \beta_1 x \\ B_2(x) &= \beta_0(1-x)^2 + 2\beta_1 x(1-x) + \beta_2 x^2 \\ B_3(x) &= \beta_0(1-x)^3 + 3\beta_1 x(1-x)^2 + \beta_2 x^2(1-x) + \beta_3 x^3 \end{aligned}$$

The first of these polynomials is just the linear function interpolating between β_0 and β_1, and in general the Bernstein polynomials of degree n should be thought of as rather roughly interpolating the coefficient sequence at the points $x = i/n$. Of course $B_\beta(0) = \beta_0$ and $B_\beta(1) = \beta_n$, but in general B_β does not take the β_k as intermediate values. As compensation, the Bernstein polynomials behave robustly with respect to variation in the constants β_k. For this reason, the Bernstein polynomials of low degree are used in computer graphics, where they are called **Bézier functions** (after a car designer who used them for practical applications). Another reason they are used in computer graphics is that they can be plotted very efficiently by means of efficient recursion properties. The intermediate values β_k for $1 \le k \le n-1$ are called in computer graphics the **control values** of the Bernstein polynomial. The figure below shows a cubic Bernstein graph with control values $(0, 1/2, 1/2, 0)$. In computer graphics, this is called a **Bézier cubic**. The control points of Bézier lines and quadratic curves have a simple geometric significance, but for higher degree curves this significance is somewhat lost. One simple characteristic, however, is that the graph of the function is always contained in the convex hull of the control points $(i/n, \beta_i)$.



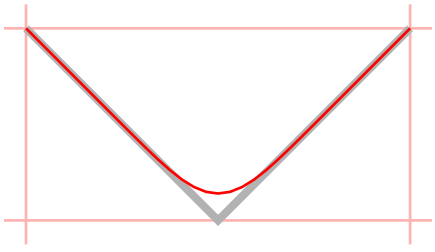
In spite of the apparent crudeness of the approximation, as n increases the Bernstein polynomials associated to f do converge to it:

2.1. Proposition. *Let f be any continuous function on $[0,1]$, and for each $n \ge 0$ let*

$$B_{f,n}(x) = \sum_{k=0}^{n} f(k/n) \binom{n}{k} x^k (1-x)^{n-k} .$$

Then the $B_{f,n}$ approach f uniformly as $n \to \infty$.

Here, for example, is the Bernstein approximation to $|x - 1/2|$ with $n = 32$.



the approximation looks good except around $x = 1/2$. It can be shown that the convergence to the value at $1/2$ is of order $1/\sqrt{n}$, which is not very good. In fact, it is not at all good idea to use Bernstein polynomals for practical approximation, in spite of theoretical virtues.

Proof. I'll follow Bernstein's original argument, with a few enhancements.

For a fixed x, the coefficients

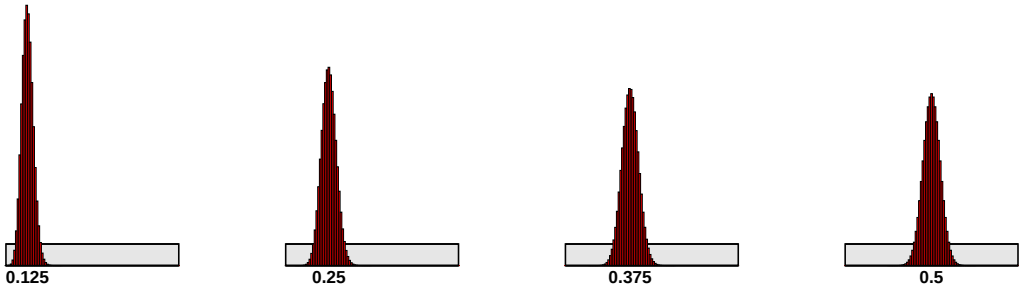
$$P_x(k/n) = \binom{n}{k} x^k (1-x)^{n-k}$$

describe the Bernoulli probability distribution assigning the probability of k successes in n independent events, each with probability x. The mean of this distribution is nx and the variance is $nx(1-x)$, and of

course for a fixed x as n grows the probability clusters around the mean, with a rough spread of $\sqrt{nx(1-x)}$ (the standard deviation). The total probability sums to 1, so in effect the distributions $P_x(k/n)$, which can be expressed as a sum of Dirac distributions

$$P_x = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \delta_{k/n} \,,$$

converge to the Dirac delta δ_x as $n \to \infty$. The Bernstein polynomial $B_{f,n}(x)$ calculates the expected value of the continuous function $f(x)$ with respect to this probability distribution, and because of the clustering of k/n around x the polynomial it should not be too surprising that $B_{f,n}(x)$ has $f(x)$ as limiting value. What is slightly subtle is that the convergence is uniform in p, but this is for simple reasons. The spread, or standard deviation, is a maximum when $p = 1/2$ and decreases to 0 at either $p = 0$ or $p = 1$, so in fact the convergence should be better near the endpoints of the interval. This is shown in the following figures, which portray the distribution for different values of x, in which $n = 100$. (These are scaled. The gray region in each is the unit square, of area 1.)



Bernstein's proof makes this intuitive reasoning rigourous. It depends on the Chebyshev inequality, which asserts that for a probability distribution with mean μ and standard deviation σ, $P(|x-\mu|) > t\sigma \leq 1/t^2$. It is only a crude estimate for Bernoulli distributions, so we should not expect the proof to give us a good estimate of convergence achieved.

Given $\varepsilon > 0$ we want to find N such that $|f(x) - B_{f,n}(x)| < \varepsilon$ for all $n \geq N$ and $0 \leq x \leq 1$. Since $[0,1]$ is compact, the function f is *uniformly continuous*—we can find $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon/2$ whenever $|x-y| < 2\delta$.

For a fixed x, the discrete variable k/n has mean value x and standard deviation $\sqrt{x(1-x)/n}$. Therefore by Chebyshev's inequality the probability P of $|x - k/n| > t\sqrt{x(1-x)/n}$ is at most $1/t^2$. Let M be the maximum spread of f across the interval $[0,1]$, that is to say the difference between its maximum and its minimum. Then

$$\begin{aligned} 1 &= \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \\ f(x) &= f(x) \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \\ &= \sum_{k=0}^{n} f(x) \binom{n}{k} x^k (1-x)^{n-k} \,. \end{aligned}$$

Therefore

$$\begin{aligned}\left|B_{f,n}(x)-f(x)\right| &\le \sum_{k=0}^{n}\left|f(k/n)-f(x)\right|\binom{n}{k}x^k(1-x)^{n-k}\\ &= \sum_{|k/n-x|\le\delta}\left|f(k/n)-f(x)\right|\binom{n}{k}x^k(1-x)^{n-k}\\ &\qquad + \sum_{|k/n-x|>\delta}\left|f(k/n)-f(x)\right|\binom{n}{k}x^k(1-x)^{n-k}\\ &\le \varepsilon/2 + M/t^2\end{aligned}$$

if $t\sqrt{x(1-x)/n}=\delta$. For all n such that $M/t^2<\varepsilon/2$ or $n>2Mx(1-x)/\varepsilon\delta^2$, we have

$$|B_{f,n}(x)-f(x)|<\varepsilon\,.\ \blacksquare$$

For completeness, I include here the proof of Chebyshev's inequality. Given a discrete probability distribution p_i with mean μ and standard deviation σ, we want to show that

$$\sum_{|x-\mu|/s\sigma\ge 1} p_i \le \frac{1}{s^2}\,.$$

But this sum is

$$\begin{aligned}\sum_{|x_i-\mu|/s\sigma\ge 1} p_i &= \sum_{|x_i-\mu|^2/s^2\sigma^2\ge 1} p_i\\ &\le \sum_{|x_i-\mu|^2/s^2\sigma^2\ge 1} p_i\frac{|x_i-\mu|^2}{s^2\sigma^2}\\ &le \sum_i p_i\frac{|x_i-\mu|^2}{s^2\sigma^2}\\ &= \frac{1}{s^2\sigma^2}\sum_i p_i|x_i-\mu|^2\\ &= \frac{1}{s^2}\,.\end{aligned}$$

3. The Stone-Weierstrass Theorem

If X is a topological space and R a subring of $C(X)=C(X,\mathbb{R})$, it is said to **separate points** of X if for every $x\ne y$ in X there exists φ in R with $\varphi(x)\ne\varphi(y)$.

3.1. Theorem. *If X is a compact Hausdorff space and R is a subring of $C(X)$ that (a) contains the constants $\mathbb{R}$ and (b) separates points of X then it is dense in $C(X)$.*

[Kelley:1955], whom I follow loosely, says of this result that it is "*unquestionably the most useful known result on $C(X)$.*"

Proof. The first step is to show that if f is in R then $|f|$ can be approximated by elements in R. Suppose $|f|$ is bounded by M on X. By Weierstrass' theorem for intervals of $\mathbb{R}$, we can find an approximation of $|x|$ in $[-M,M]$ by polynomials $P(x)$, hence an approximation of $|f|$ by $P(f)$, which lies in the ring R.

Now

$$\begin{aligned}\max(f,0) &= \frac{f+|f|}{2}\\ \max(f-g,0) &= \frac{(f-g)+|f-g|}{2}\\ \max(f,g) &= f-\max(f-g,0)\\ &= \frac{(f+g)+|f-g|}{2}\,.\end{aligned}$$

It follows that if f and g are two functions in R, both $\min(f,g)$ and $\max(f,g)$ are in $\overline{R}$.

I now follow [Pinkus:2000]. Since R separates points and contains the constants, given any $x \neq y$ in X we can find a function φ in R with $\varphi(x) \neq \varphi(y)$. The function

$$h(w) = \frac{\varphi(w)-\varphi(x)}{\varphi(y)-\varphi(x)}$$

is in R and satisfies $h(x)=0$, $h(y)=1$, so functions in R can interpolate any two values on X.

Suppose f in $C(X)$, $\varepsilon > 0$. Fix temporarily x in X. For any y we can find h in R such that $h(x)=f(x)$, $h(y)=f(y)$, and we can then find a neighbourhood U_y of y such that $|h(w)-f(w)|<\varepsilon$ for all w in U_y. In particular $h(w) > f(w)-\varepsilon$ for all w in U_y. Choose a finite subcover, so now we are given a collection of open sets U_i and functions h_i such that for all i (a) $h_i(x)=f(x)$ and (b) $h_i(w) > f(w)-\varepsilon$ on U_i. If H_x is the maximum of these then (a) $H_x(x)=f(x)$ and (b) $H_x(w) > f(w)-\varepsilon$ for all w.

To conclude, make a similar argument about neighbourhoods of each x, and take a minimum of a finite collection of functions. ∎

3.2. Corollary. *If X is a compact subset of $\mathbb{R}^n$, then every continuous function on X may be approximated arbitrary closely by polynomials in the coordinates.*

3.3. Corollary. *If X is a compact subset of $\mathbb{C}^n$, then every continuous function on X may be approximated arbitrary closely by polynomials in the coordinates and their complex conjugates.*

4. References

1. Sergei Bernstein, 'Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités', *Comm. Soc. Math. Kharkov* **13** (1912/13), 1–2. This and other classic papers in approximation theory can be found at Allan Pinkus' very useful web site

`http://www.math.technion.ac.il/hat/papers.html`

2. Peter Henrici, **Essentials of numerical analysis with pocket calculator demonstrations**, Wiley, 1982.

3. John Kelley, **General Topology**, Van Nostrand, 1955.

4. Allan Pinkus, 'Weierstrass and Approximation Theory', *Journal of Approximation Theory* **107** (2000), 1–66. Also available at

`http://www.math.technion.ac.il/hat/articles.html`