Reference:

1. Chapter 4.4 and 4.9 in [CFZ09].
2. Barron's original paper and followup, [Bar93], [Bar94].

1 The curse of dimensionality

In statistics and machine learning, the standard setting is the following. Given $(x_i \in B, y_i)$, a function $f_\theta(x)$, find parameters $\widetilde{\theta}$ such that $f_{\widetilde{\theta}}(x_i) \approx y_i$. What is our metric of success? We want an algorithm that minimizes the **mean integrated square error** (let $\widetilde{f} = f_{\widetilde{\theta}}$)

$$\text{MISE} = \mathbb{E}_{(x_i,y_i)}[\left\| \widetilde{f} - f \right\|_2^2]$$

where $\|f\|_2 = \int_B f(x)^2 \, d\mu(x)$. The expected value is over independent random (x_i, y_i), and the integral is with respect to the probability distribution on the samples x_i. (We assume there is a distribution on the x's.)

The **curse of dimensionality** is a general phenomenon where estimates degrade with the number of dimensions. Consider a class of models $f_{p,\theta} : \mathbb{R}^p \to \mathbb{R}$. For useful classes of models, the MISE typically increases superlinearly in p, and the number of data points required also increases rapidly.

Now let's look at the setting of neural nets.

Definition 1.1: A sigmoidal function is a differentiable function f on $\mathbb{R}$ with $f' > 0$, $\lim_{x\to-\infty} f(x) = 0$, and $\lim_{x\to\infty} f(x) = 1$.

We have the following.

Proposition 1.2: Let ϕ be sigmoidal. Every continuous function on a bounded set $B \subseteq \mathbb{R}^p$ can be approximated to arbitrary precision by a linear combination of $\phi(a \cdot x + b)$.

Such a combination is represented by a (1-layer) neural net where

- the input layer has p nodes, i.e., represents an element of $\mathbb{R}^p$,
- the hidden layer has some number of nodes,
- the output node is a linear combination of hidden layer nodes.

(We're trying to approximate a function rather than make a decision, so we don't take a threshold function at the output.)

A natural question is how well can such a neural net approximate an arbitrary continuous function? We'll give a precise answer, depending on the regularity of f and the size of the hidden layer we allow, but *not the dimension p*. Barron's theorem tells us that "neural nets evade the curse of dimensionality" in the following sense.

The best 1-layer neural net approximations do not get worse as p increases.

Since the number of hidden nodes required to get ε approximation does not depend on p, and the number of parameters is $O((\text{hidden nodes})\, p)$, the number of parameters we need is linear in p rather than superlinear.

Note we are not saying anything about an *algorithm* to find the best approximation. The loss function is in general not convex so it's unclear whether gradient descent will actually find the approximation that Barron's Theorem gives. If we care about the actual $\widetilde{f}$ that we estimate from sample data (x_i, y_i), then we have to include the estimation error:

$$\text{MISE} \le \|f - f_{\text{best}}\|_2 + \mathbb{E}_{(x_i,y_i)} \left\| f_{\text{best}} - \widetilde{f} \right\|_2 .$$

We'll discuss the first term in the Section 2, and the second term in Section **??**.

2 Barron's Theorem

The Fourier transform of $f : \mathbb{R}^p \to \mathbb{R}$ is

$$\widehat{f}(\omega) = \frac{1}{(2\pi)^p} \int_{\mathbb{R}^p} f(x) e^{-i\omega\cdot x}\, dx.$$

The Fourier inversion formula is

$$f(x) = \int \widehat{f}(x) e^{i\omega\cdot x}\, dx.$$

When B is the unit ball, our measure of smoothness will be the following:

$$\left\| \widehat{f'} \right\|_1 = \left\| \omega \widehat{f} \right\|_1 = \int_{\mathbb{R}^p} |\omega \widehat{f}(\omega)|\, d\omega.$$

More generally, for an arbitrary bounded set $B \subseteq \mathbb{R}^p$ let $|\omega|_B = \sup_{x\in B} |\omega \cdot x|$. When $B = B_0(1)$ is the unit ball, this is simply $\|\omega\|_2$. In the general setting our smoothness measure is

$$\|f\|_B^* := \int_{\mathbb{R}^p} |\omega|_B |\widehat{f}(\omega)|\, d\omega.$$

Let Γ_B be the set of functions on B where the Fourier inversion formula holds after subtracting out the mean,[1]

$$\Gamma_B = \left\{ f : B \to \mathbb{R} : \forall x \in B, f(x) = f(0) + \int (e^{i\omega\cdot x} - 1)\widehat{f}(\omega)\, d\omega \right\}$$

Let $\Gamma_{B,C}$ be the subset with smoothness $\le C$:

$$\Gamma_{B,C} = \Gamma_B \cap \{\|f\|_B^* \le C\}.$$

The quality of the approximation will depend on how large the phases of f are. We'll see in the proof where the norm $\|f\|_B^*$ arises.

[1] for example, it includes all smooth (C^∞) functions and more generally, all L^1 functions on B whose Fourier transform is also L^1

Theorem 2.1 (Barron)**:** Let $B \subseteq \mathbb{R}^p$ be a bounded set, μ a probability measure on B, and $\varepsilon > 0$. Let $f \in \Gamma_{B,C}$ and ϕ be sigmoidal. There exists

$$f_r = \sum_{i=1}^{r} c_i \phi(a_i \cdot x + b_i)$$

with $\sum_{i=1}^r |c_i| \le 2C$ such that

$$\|f - f_r\|^2 = \int_B (f(x) - f_r(x))^2 \, \mu(dx) \le \frac{(2C)^2}{r} + \varepsilon.$$

We'll just consider the case when μ is uniform on B, but in general, the proof goes through the same way (with a bit more care).

This means that the number of parameters required to get an approximation of ε is $(p+2)r = (p+2)\frac{(2C)^2}{\varepsilon}$, which is linear in p rather than superlinear.

The idea of the proof is the following.

1. Show that f is in the closed convex hull of the ϕ's. We break this into several inclusions which we show one at a time:

$$\begin{aligned}\{\|f\|^* \le C\} &\overset{(3)}{\subseteq} \overline{\text{conv}} \underbrace{\left\{\frac{\gamma}{|\omega|_B}(\cos(\omega \cdot x + b) - \cos b) : \omega \ne 0, |\gamma| \le C\right\}}_{=:G_{\cos}} \\ &\overset{(2)}{\subseteq} \overline{\text{conv}} \underbrace{\{cH(a \cdot x + b) : c \le 2C, |a|_B = 1, |b|_B \le 1\}}_{=:G_{\text{step}}} \\ &\overset{(1)}{\subseteq} \overline{\text{conv}} \underbrace{\{c\phi(a \cdot x + b) : c \le 2C\}}_{G_\phi}\end{aligned}$$

   where H is the step function $1_{x \ge 0}$. We explain the inclusions. First, the exact form of ϕ doesn't matter: all we need about ϕ is that it can approximate step functions arbitrarily well. (ϕ sigmoidal gives us this.)

2. Second, we write the step functions in terms of a standard basis, namely the Fourier basis.

3. Third, we write out the Fourier expansion of an arbitrary regular f to show that f is in $G_{\cos}$.

4. Next, we use a general fact: If A is convex and $f \in \overline{\text{conv}} A$, then f is close to a small combination of elements of A. This in fact holds in any Hilbert space. The proof is by writing f as a linear combination, and then sampling the functions with probabilities given by the coefficients.

   Thus f being in the convex hull of the ϕ's gives us that f is close to a small combination of them.

Proof. 1. Without loss of generality, ϕ is centered at 0. Then

$$\phi(k(a\cdot x+b)) \to H(a\cdot x+b)$$

for $x \neq 0$ so $G_{\text{step}} \subseteq \overline{G_\phi}$.

2. We relate H to the the Fourier basis: $G_{\cos} \subseteq \overline{\text{conv}}(G^\mu_{\text{step}})$. We can do this easily because each $\cos(\omega\cdot x+b)-\cos b$ is 1-dimensional. (This is why Fourier transforms are useful in this proof: $\omega \cdot x + b$ is a projection of x onto the ω direction.)

   Let $g(y) = \cos(|\omega|_B y + b) - \cos(y)$. Let $x_{-k}, \ldots, x_k$ be a partition of $[-1,1]$ such that g changes by $< \varepsilon$ on each interval, we can approximate g to within ε at every point by the sum

$$\sum_{i\geq 0}(g(x_{i+1})-g(x_i))1_{\geq x_i} + \sum_{i\leq 0}(g(x_{i-1})-g(x_i))1_{\leq x_i}.$$

   The sum of coefficients is

$$\sum_i |g(x_{i+1})-g(x_i)| \leq \int |g'|\,dx \leq 2|\omega|_B.$$

   Now substitute $y = \frac{\omega}{|\omega|_B}\cdot x$ to get the approximation of $\cos(\omega\cdot x + b)$ by a linear combination with sum of coefficients $2|\omega|_B$, i.e., an approximation of $\frac{\gamma}{|\omega|_B}(\cos(\omega\cdot x + b) - \cos b), \omega \neq 0, |\gamma| \leq C$ by a linear combination of H's with sum of coefficients $2C$. [2]

3. When is $f \in \overline{\text{conv}}(G_{\cos})$? We show $\{\|f\|^* \leq C\} \subseteq \overline{\text{conv}}(G_{\cos})$. Use Fourier inversion. Write the Fourier transform in polar form as $\widehat{f} = |\widehat{f}|e^{i\theta(\omega)}$:

$$\begin{aligned}
f(x) - f(0) &= \int \widehat{f}(\omega)(e^{i\omega\cdot x}-1)\,d\omega \\
&= \int |\widehat{f}|e^{i\theta(\omega)}(e^{i\omega\cdot x}-1)\,d\omega \\
&= \int |\widehat{f}|(\cos(\omega\cdot x+\theta(\omega)) - \cos(\theta(\omega)))\,d\omega && \text{taking real part} \\
&= \int |\widehat{f}||\omega|_B \frac{1}{|\omega|_B}(\cos(\omega\cdot x+\theta(\omega)) - \cos(\theta(\omega)))\,d\omega.
\end{aligned}$$

   Hence, so long as $\int |\widehat{f}||\omega|_B \leq C$, f is in a combination of functions in $G_{\cos}$ with sum (integral) of coefficients $\leq C$. (The integral is in the closure of the convex hull because it can be approximated as a Riemann sum.)

4. We show the following.

[2] For an arbitrary measure, there is an extra step where we show that we can restrict $-b$ to the continuity points of the measure μ.

Lemma 2.2: Let G be a bounded set in a Hilbert space, where every element has norm $\le b$. (For example, $G \subseteq L^2(B)$.) Let $f \in \overline{\text{conv}}(G)$. Then for every r,

$$\inf_{f_r=\sum_{i=1}^r c_i g_i, g_i \in G, \sum c_i = 1} \|f - f_r\|^2 \le \frac{b^2 - \|f\|^2}{r} \le \frac{b^2}{r}.$$

(The infimum is taken over all convex combinations involving r functions.)

Proof. Since $f \in \overline{\text{conv}}(G)$, for all ε, there exists f^* in the following form that is ε away from f:

$$f \approx_\varepsilon f^* = \sum_{i=1}^m c_i g_i^*.$$

Let g be a random variable such that

$$g = g_i^* \text{ with probability } \frac{c_i}{\sum_{j=1}^m |c_j|}.$$

Let $g_1, \dots, g_r$ be r independent draws, and let f_n be the average,

$$f_r = \frac{1}{r}\sum_{i=1}^r g_i.$$

Then (since f_r is the average of r variables distributed as G and $f^* = \mathbb{E}g$)

$$\begin{aligned}\mathbb{E}\|f_r - f^*\|^2 &= \frac{1}{r}\mathbb{E}\|g - \mathbb{E}g\|^2 \\ &= \frac{1}{r}[\mathbb{E}(g^2) - (\mathbb{E}g)^2] \\ &\le \frac{1}{r}(b^2 - \|f\|^2).\end{aligned}$$

□

Finally, apply the lemma to

$$f \in \overline{\text{conv}}\left\{c\phi(a \cdot x + b) : c \le 2C\right\},$$

noting that the norms of the ϕ's are ≤ 1 since μ is a probability measure.

□

3 Barron's Theorem: Estimation error

We'll be very sketchy in this section.

Theorem 3.1: Let $\widehat{f}_{r,N,C}(x)$ be the least squares neural net estimator, calculated from N samples (x_i, y_i). Let the true function have $\|f\|^* \le C$. Suppose that $\left\|\phi\left(\frac{x}{\tau(\varepsilon)}\right) - 1_{\ge 0}\right\|_2 \le \varepsilon$ for $\tau(\varepsilon)$ polynomial in ε. Then

$$\text{MISE} = \mathbb{E}_{(x_i,y_i)}\left\|f - \widetilde{f}_{r,N,C}\right\|^2 \le O\left(\frac{C^2}{r}\right) + O\left(\frac{rp}{N}\ln N\right).$$

The first ter, becomes smaller for r large because more hidden nodes improves the best possible approximation. The second term becomes larger for r large because the estimate degrades due to overfitting. Balance these terms by setting $n \sim \|f\|^* \left(\frac{N}{p\ln N}\right)^{\frac{1}{2}}$; we get error $O(\|f\|^* \left(\frac{p}{N}\ln N\right)^{\frac{1}{2}})$. Note two things.

1. The exponent $\frac{1}{2}$ is independent of d.
2. Thus we can take $N = \widetilde{O}(kp)$ to get a fixed approximation up to a fixed ε.

In many other models (fitting to polynomials, etc.), the error goes like $\left(\frac{1}{N}\right)^{\frac{C}{p}}$ or $\left(\frac{1}{N}\right)^{\frac{2s}{2s+p}}$ for some smoothness parameter s. For fixed ε, this requires the number of samples to be $\left(\frac{1}{\varepsilon}\right)^{\frac{p}{C}}$, exponential in the dimension.

Note, however, that there isn't a very good way to find the $\widehat{f}$ other than brute-force trying all the parameters in an ε-net, which is not an efficient algorithm.

Proof sketch.

1. Let $\Theta_{r,\varepsilon,\tau,C}$ be an ε-net for the parameters. Letting $\Theta_{r,\tau,C}$ be the set of parameters with $|a_k|_1, |b_k|_1 \le \tau, \sum |c_k| \le C$, we choose $\Theta_{r,\varepsilon,\tau,C}$ such that for every θ there exists $\theta^* \in \Theta_{r,\varepsilon,\tau,C}$ with $|a_k - a_k^*|, |b_k - b_k|^* \le \varepsilon, \sum |c_k - c_k^*| \le C\varepsilon, |c_0 - c_0^*| \le C\varepsilon$. We can $\Theta_{r,\varepsilon,\tau,C}$ having size

$$\ln \Theta_{r,\varepsilon,\tau,C} \le \underbrace{(r(p+2)+1)}_{\#\text{ parameters}} \underbrace{\ln\left(\frac{\text{poly}(r)}{\varepsilon}\right)}_{\text{grid width}}.$$

2. By a PAC-learning type bound (with least squares error rather than ERM—expected regret minimization for classification), we get

$$\left\|f_{\text{best}} - \widetilde{f}\right\| \le O(\frac{rp}{N}\ln N).$$

Here the estimated parameters are

$$\widetilde{\theta}_{r,N} = \text{argmin}_{\theta \in \Theta_N}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - f_r(x_i,\theta))^2\right).$$

□

References

[Bar93] Andrew R. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". In: *IEEE Transactions on Information Theory* 39.3 (1993), pp. 930–945. ISSN: 00189448. DOI: 10.1109/18.256500.

[Bar94] Andrew R. Barron. "Approximation and estimation bounds for artificial neural networks". In: *Machine Learning* 14.1 (1994), pp. 115–133. ISSN: 08856125. DOI: 10.1007/BF00993164.

[CFZ09] Bertrand Clarke, Ernest Fokoue, and Hao Helen Zhang. *Principles and Theory for Data Mining and Machine Learning*. Vol. 26. 2003. 2009, pp. 251–264. ISBN: 9780387981345. DOI: 10.1007/978-0-387-98135-2. URL: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-98134-5?cm%5C_mmc=AD-%5C_-Enews-%5C_-ECS12245%5C_V1-%5C_-978-0-387-98134-5.