

Overfitting & Generalization

Mathematics for Machine Learning

Lecturer: Matthew Wicker

A few notes:

- Lecture 4 and 5 make ups will be posted today
- Monday after next is when the coursework is due
- Final lecture schedule/topics:
 - Overfitting & Generalization
 - More Concentration & Cross-Validation
 - Bias-Variance Trade-off
 - PCA
- Practice exam to be posted on Scientia with equation sheet following PCA lecture

Material Covered

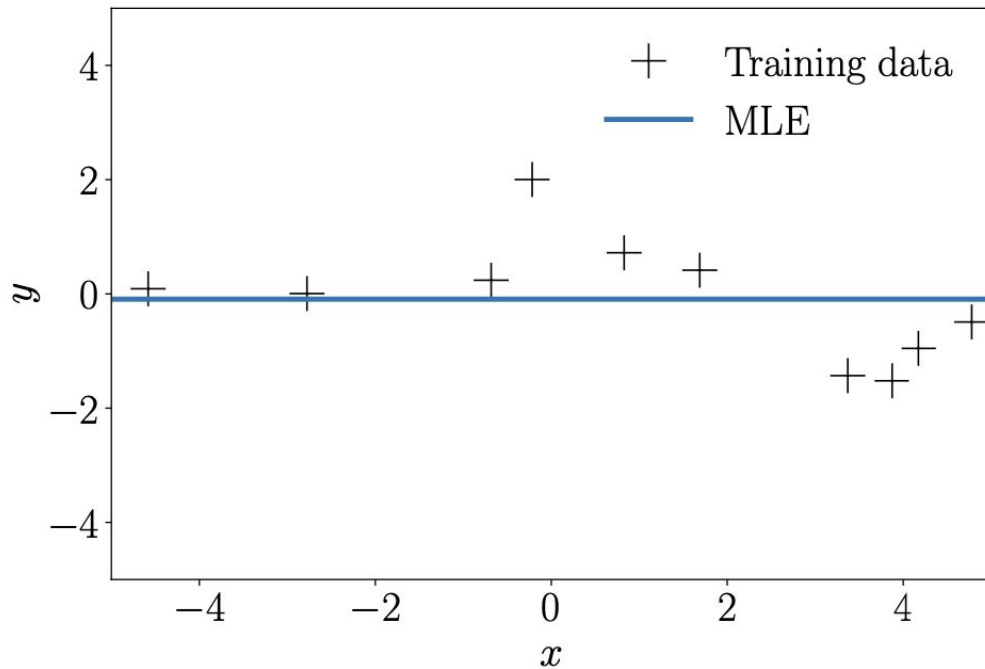
Models: Linear models, basis expansion, logistic regression, neural networks, Prob. densities, Bayesian density estimation

Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity, Maximum likelihood, maximum a posteriori, Bayesian inference, LOTUS, change of variables, expectation identities, equating coefficients, joint Gaussian, epistemic/aleatoric uncertainty

Settings: Regression, Classification, Density Estimation

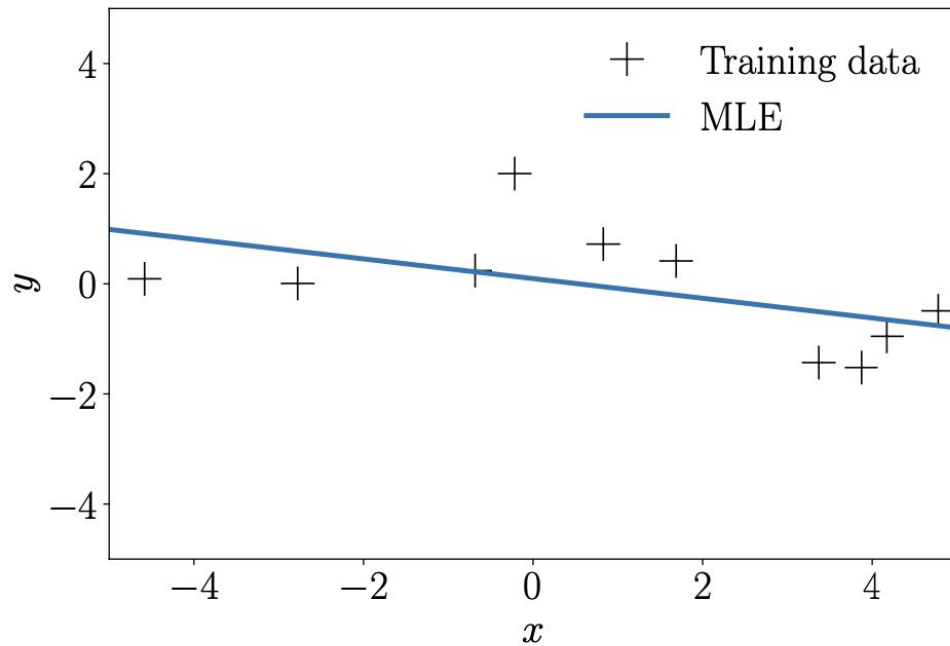
This lecture: Universal function approximation, overfitting, test set, concentration inequalities

Fitting different linear models



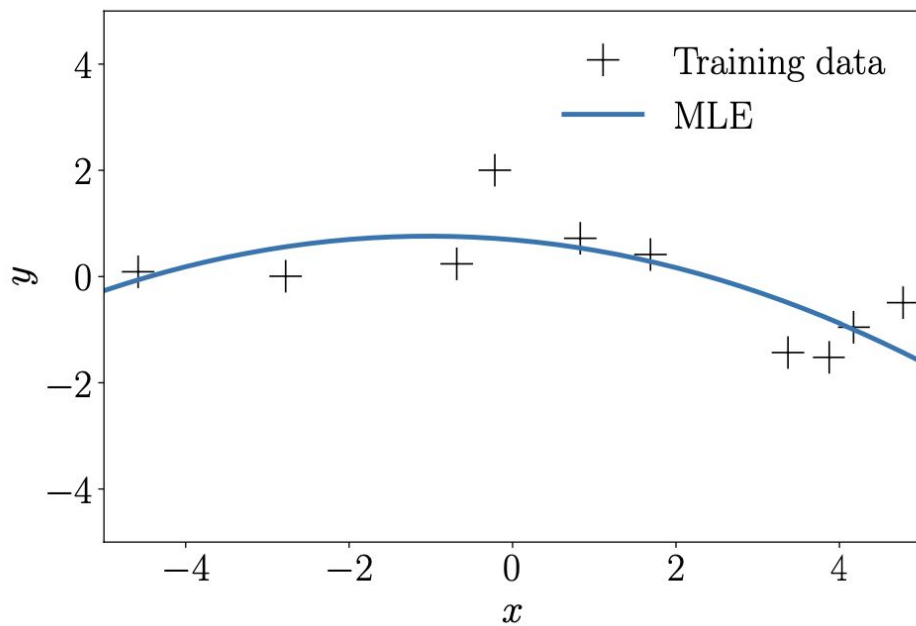
$$\phi(x) = [1]^T$$

Fitting different linear models



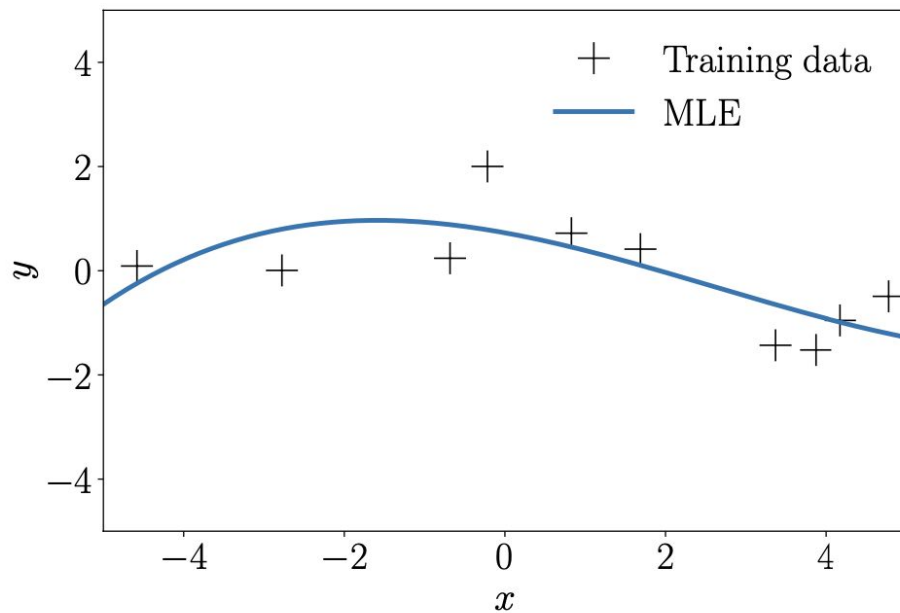
$$\phi(x) = [1 \ x]^T$$

Fitting different linear models



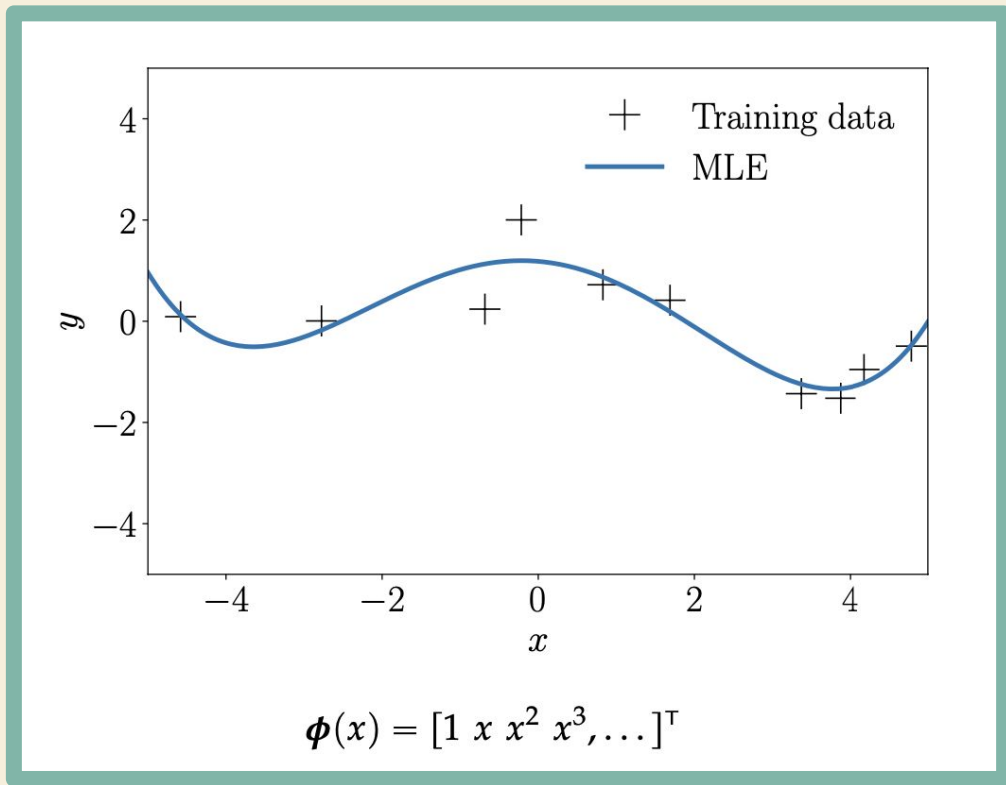
$$\phi(x) = [1 \ x \ x^2]^\top$$

Fitting different linear models

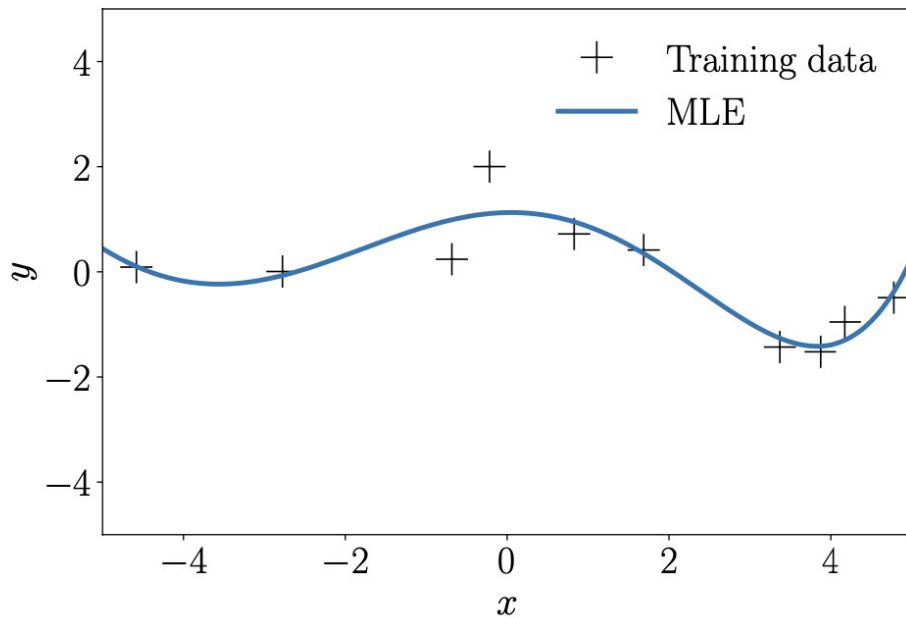


$$\phi(x) = [1 \ x \ x^2 \ x^3]^\top$$

Fitting different linear models

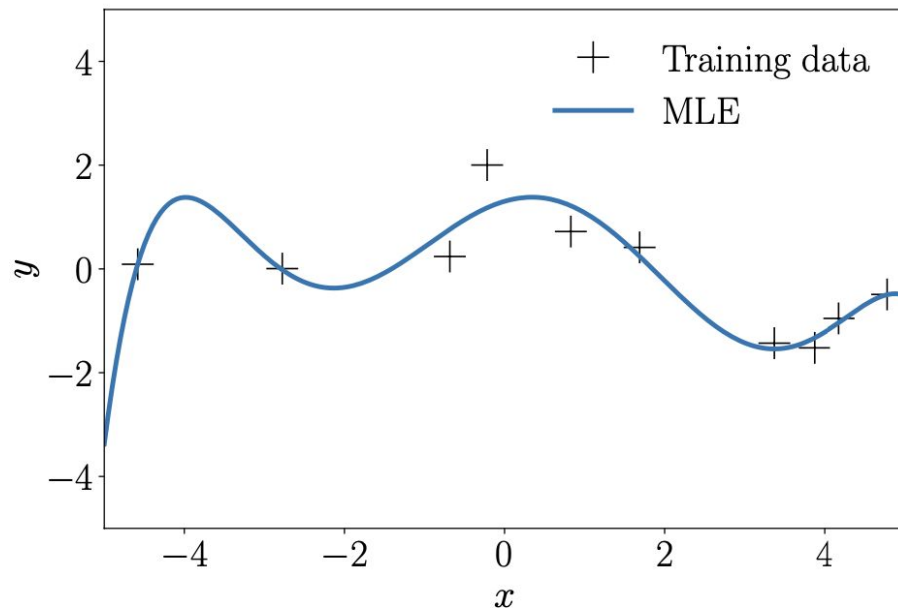


Fitting different linear models



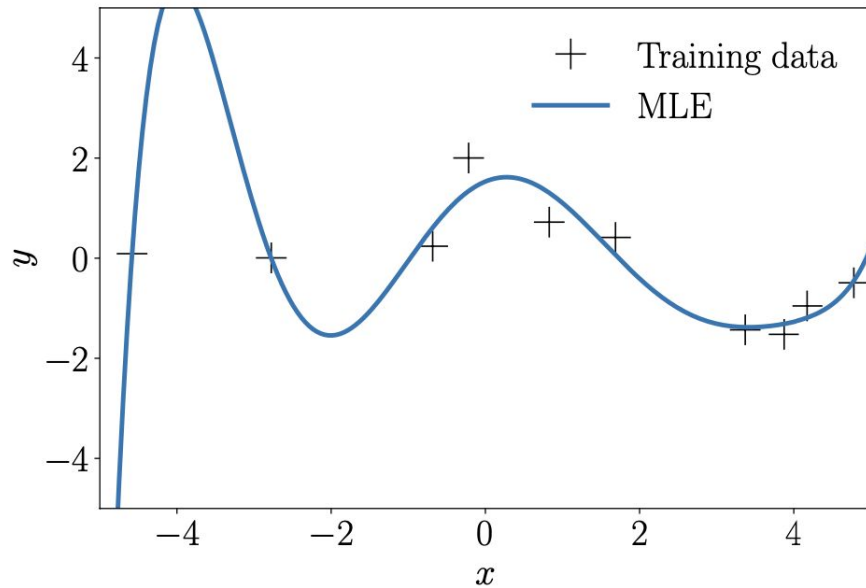
$$\phi(x) = [1 \ x \ x^2 \ x^3, \dots]^\top$$

Fitting different linear models



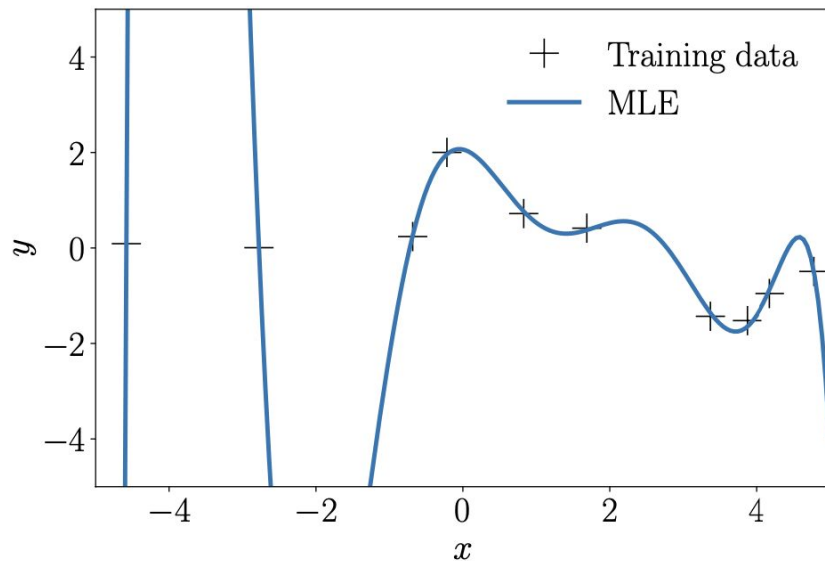
$$\phi(x) = [1 \ x \ x^2 \ x^3, \dots]^\top$$

Fitting different linear models



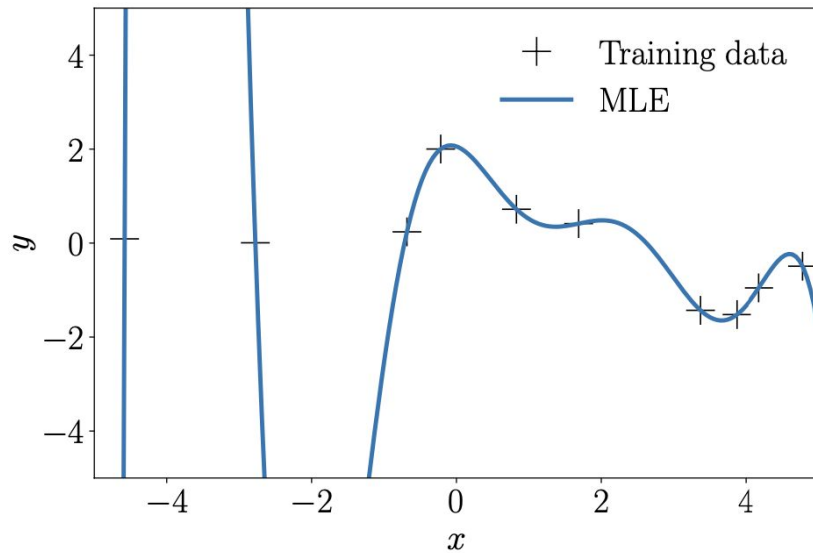
$$\phi(x) = [1 \ x \ x^2 \ x^3, \dots]^\top$$

Fitting different linear models



$$\phi(x) = [1 \ x \ x^2 \ x^3, \dots]^\top$$

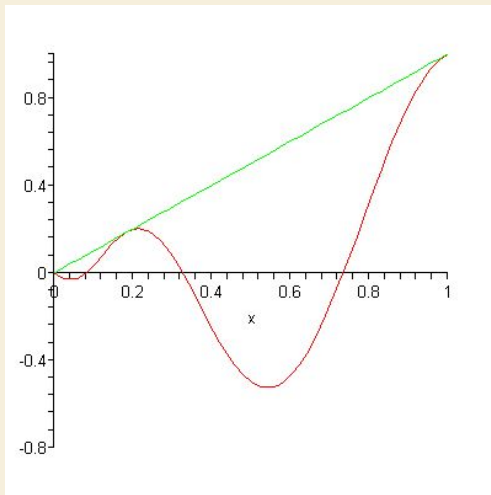
What is happening to our loss?



$$\phi(x) = [1 \ x \ x^2 \ x^3, \dots]^\top$$

Stone-Weierstrass Theorem

Bernstein polynomials:



There is a polynomial to model data

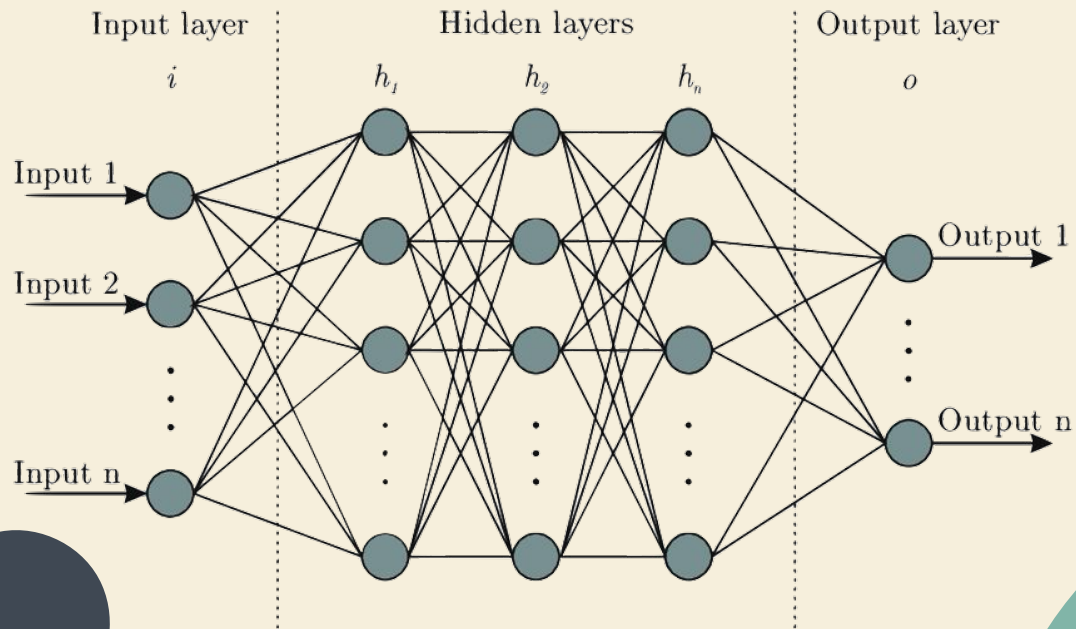
Definition 2.1. Weierstrass Theorem Let f be a continuous function on a closed interval $[a, b]$. For any $\varepsilon > 0$, there exists a polynomial function $P(x)$ such that

$$\|f - P\|_{\infty} = \sup_{x \in [a, b]} |f(x) - P(x)| < \varepsilon.$$

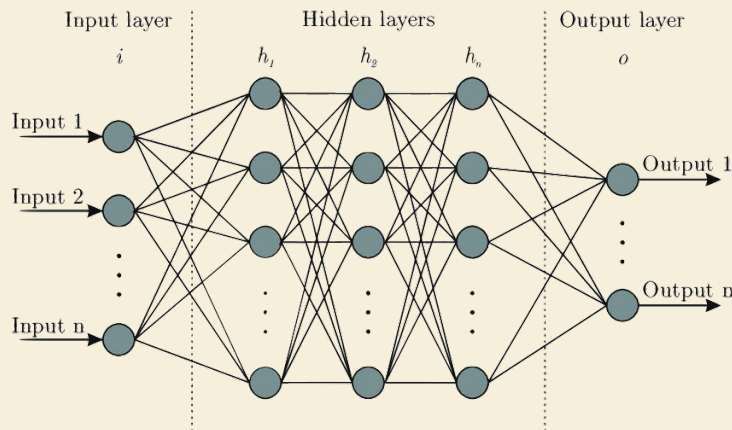
↙ arbitrary value

In other words, the polynomial $P(x)$ can uniformly approximate f on the interval $[a, b]$.

Do other models have this?



Neural Networks



Definition 2.2. Universal Approximation Theorem (Cybenko, 1989) Let $\sigma(x)$ be a sigmoidal activation function, i.e., a function that is nonconstant, bounded, and continuously differentiable. For any continuous function f on a compact subset of \mathbb{R}^n , and any $\varepsilon > 0$, there exist positive integers N , weights w_i , and biases b_i , and a sum of N sigmoidal functions such that:

$$F(x) = \sum_{i=1}^N w_i \sigma(w_i^T x + b_i)$$

satisfies $|F(x) - f(x)| < \varepsilon$ for all x in the compact subset.

A more general statement

Theorem 2.3 (Stone-Weierstrauss Theorem (limited version)). *Let F be a class of functions defined on a compact set $S \subseteq \mathbb{R}^d$. If F satisfies:*

- 1. Each $f \in F$ is continuous.*
- 2. For every x , there exists $f \in F$ such that $f(x) \neq 0$.*
- 3. For every x, x_1 with $x \neq x_1$, there exists $f \in F$ such that $f(x) \neq f(x_1)$ (F separates points).*
- 4. F is closed under multiplication ($\forall f, g \in F, \exists h \in F$ such that $h(x) = f(x)g(x)$) and vector space operations (F is an algebra).*

Then, for every continuous function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ and any $\epsilon > 0$, there exists $f \in F$ such that $\|f - g\|_\infty \leq \epsilon$. In other words, F is a universal approximator.

What have we said?

Given the data-generating function is a continuous function
We can always fit that continuous function with our ML models

We have not said:

Fitting a finite amount of data implies fitting the data-generating function of interest

(fitting finite amount of data \rightarrow VC dimension/shattering)

Picking a model at random

Imagine we have selected one of the previous models we just discussed completely at random. We train it, and, as expected, the loss function is very low. We can measure how well it does after deployed:

$$\text{Err}_{\text{deploy}}(\mathbf{x}, y^{(i)}) = \ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)}))$$

eg. $\|y - \theta\|_2^2$

Picking a model at random

Measuring how well a model performs over a trial run:

$$\text{Err}_{\text{deploy}} = \frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)}))$$

Picking a model at random

If this trial run is infinitely long then we end up with the expectation over the joint distribution:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^\theta(\mathbf{x}^{(i)})) = \overset{\substack{\text{expectation of} \\ \downarrow \text{loss}}}{\mathbb{E}_{p(\mathbf{x}, y)}} \left[\ell(y^{(i)}, f^\theta(\mathbf{x}^{(i)})) \right]$$

How is this different from our training loss?

Held out test set

$$\mathbb{E}_{p(\mathbf{x}, y)} \left[\ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)})) \right]$$

- Measure the "generalization" after training
- Measure more than just the loss: consider harms that can come from our model
- The error function may not be the same as our loss: cross-entropy and 0-1 error.

Dealing with only finite data^a

We do not have access to an infinitely long trial, unfortunately

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^\theta(\mathbf{x}^{(i)})) = \mathbb{E}_{p(\mathbf{x}, y)} \left[\ell(y^{(i)}, f^\theta(\mathbf{x}^{(i)})) \right]$$

So what can we do? One option is to compute the integral over the space of data. Why can we rule that out?

Held out test set

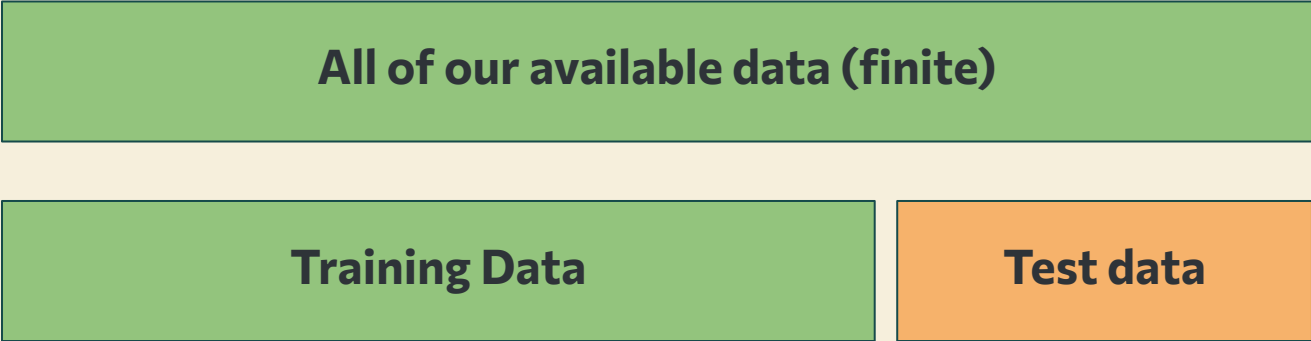
$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)})) = \mathbb{E}_{p(\mathbf{x}, y)} \left[\ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)})) \right]$$

All of our available data (finite)

Training Data

Test data

We estimate our error at
deploy time with the test data



Held out test set

$$\frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)})) \approx \mathbb{E}_{p(\mathbf{x}, y)} \left[\ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)})) \right]$$

All of our available data (finite)

Training Data

Test data

We estimate our error at
deploy time with the test data



In what sense is this estimator good?

$$\frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)})) \approx \mathbb{E}_{p(\mathbf{x}, y)} \left[\ell(y^{(i)}, f^{\theta}(\mathbf{x}^{(i)})) \right]$$

Training Data

Test data

proportion?

Training Data

Test data

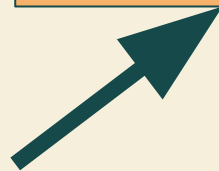
Training Data

Test
data

What is the mathematical object?

Training Data

Test data



$$\{(X_1 = x_1, Y_1 = y_1), (X_2 = x_2, Y_2 = y_2), \dots, (X_N = x_N, Y_N = y_N)\}$$

$$\{Z_1 = \ell(f^\theta, X_1 = x_1, Y_1 = y_1), \dots, Z_N = \ell(f^\theta, X_N = x_N, Y_N = y_N)\}$$

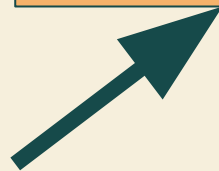
What assumptions do we need to make here?

Joint loss $x, y \rightarrow z$

What is the mathematical object?

Training Data

Test data



$$\{(X_1 = x_1, Y_1 = y_1), (X_2 = x_2, Y_2 = y_2), \dots, (X_N = x_N, Y_N = y_N)\}$$

$$\{Z_1 = \ell(f^\theta, X_1 = x_1, Y_1 = y_1), \dots, Z_N = \ell(f^\theta, X_N = x_N, Y_N = y_N)\}$$

$$\text{avg} \rightarrow \hat{Z} = \frac{1}{N} \sum_{i=1}^N Z_i = \ell(f^\theta, x_i = x_i, y_i = y_i)$$

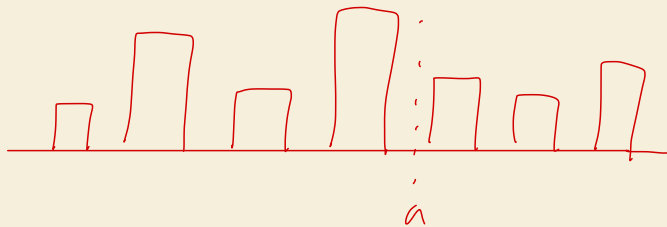
Markov's inequality

Assume Z is discrete & non -ve

$$Z \geq 0 \quad \mathbb{E}[Z] = \sum_z z P(Z = z)$$

$$\sum_z z P(Z = z) \geq \sum_{z > a} z P(Z = z)$$

$z > a$ ← just a number



$E(Z)$ for all points \geq , $E(Z)$ for points past a

Markov's inequality

$$Z \geq 0$$

$$\mathbb{E}[Z] = \sum_z z P(Z = z)$$

$$\sum_z z P(Z = z) \geq \sum_{z > a} z P(Z = z) \geq \sum_{z > a} a P(Z = z)$$

now const -

z is increasing with sum, a const

Markov's Inequality

$$Z \geq 0$$

$$\mathbb{E}[Z] = \sum_z z P(Z = z)$$

$$\boxed{\mathbb{E}[Z]} = \sum_z z P(Z = z) \geq \sum_{z \geq a} z P(Z = z) \geq \sum_{z \geq a} a P(Z = z) = \boxed{a P(Z > a)}$$

$\mathbb{E}[X] = \sum_x x P(X=x) \geq \sum_{x \geq a} x P(X=x) \geq \sum_{x \geq a} a P(X=x) = a P(X \geq a)$

$= a \sum_{z \geq a} P(Z=z)$

We have related the expectation to the probability:

If the expectation is small, then the probability ^{that} Z is large is small

Using Markov's inequality

$$\mathbb{E}[Z] \geq aP(Z > a)$$

$$\mathbb{E}[(Z - \mu)^2] \geq a^2 P((Z - \mu)^2 \geq a^2)$$

Using Markov's inequality

$$\mathbb{V}[Z] = \mathbb{E}[(Z - \mu)^2]$$

$$\geq a^2 P((z - \mu)^2 \geq a^2) = a^2 P(|z - \mu| \geq a)$$

We have now related the variance to a probability in a similar way. This inequality says that if the variance is small, so is the probability that your sample is far from the mean

Using Markov's inequality

$$\mathbb{V}[Z] = \mathbb{E}[(Z - \mu)^2]$$

$$\geq a^2 P((z - \mu)^2 \geq a^2) = a^2 P(|z - \mu| \geq a)$$

Importantly, we have made an implicit assumption about the random variable having finite mean and variance

Using Markov's inequality

$$\sigma^2 \geq a^2 P(|z - \mu| \geq a)$$

$$\frac{\sigma^2}{a^2} \geq P(|z - \mu| \geq a)$$

Chebyshev's inequality

$$\sigma^2 \geq a^2 P(|z - \mu| \geq a)$$

$$\frac{\sigma^2}{a^2} \geq P(|z - \mu| \geq a)$$

Plug in $a = k\sigma$ and we get Chebyshev's inequality
k x sigma

$$\frac{1}{k^2} \geq P(|z - \mu| \geq k\sigma)$$

Returning to our testing case

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^\theta(\mathbf{x}^{(i)})) = \mathbb{E}_{p(\mathbf{x}, y)} [\ell(y^{(i)}, f^\theta(\mathbf{x}^{(i)}))]$$

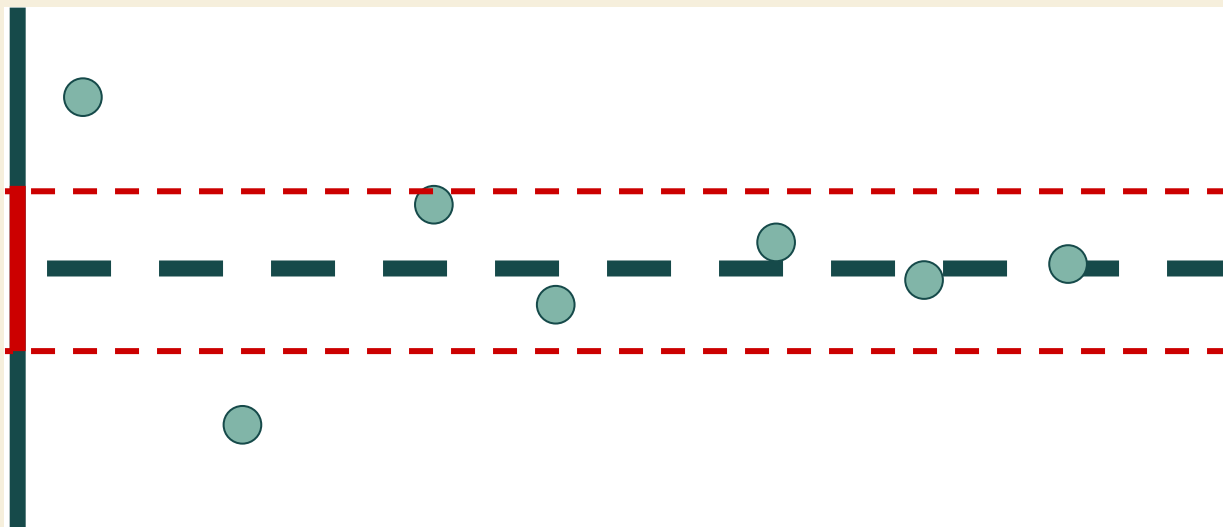
Recall that we want to reason about if the limit of our error estimate *converges* to our expectation.

Definition 2.1. Convergence A series x_1, x_2, \dots, x_n is said to *converge* to a limit L if for any $\epsilon > 0$ we have an integer K such that $\forall M > K, |x_M - L| < \epsilon$.

gets arbitrarily close to limit after a certain point

Convergence of sum of rv

Definition 2.1. Convergence A series x_1, x_2, \dots, x_n is said to *converge* to a limit L if for any $\epsilon > 0$ we have an integer K such that $\forall M > K, |x_M - L| < \epsilon$.



series falls
within range
and stays
there

Returning to our testing case

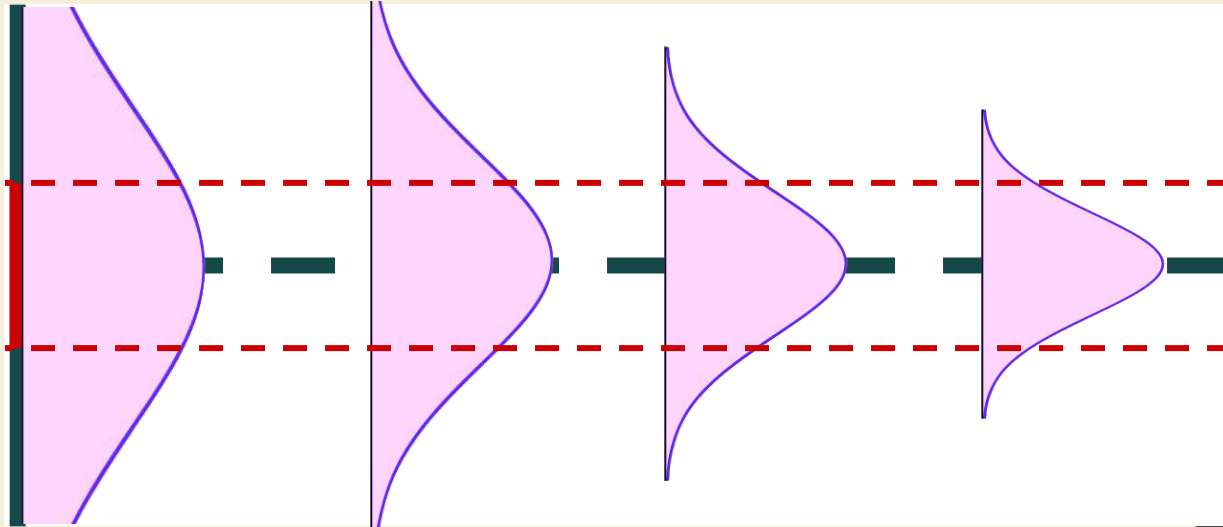
$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=0}^N \ell(y^{(i)}, f^\theta(\mathbf{x}^{(i)}))$$

$$\{Z_1 = \ell(f^\theta, X_1 = x_1, Y_1 = y_1), \dots, Z_N = \ell(f^\theta, X_N = x_N, Y_N = y_N)\}$$

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^N Z_i$$

Recall that we want to reason about if the limit of our error estimate converges to our expectation.
But we have a sum of random variables

Convergence of sum of rv

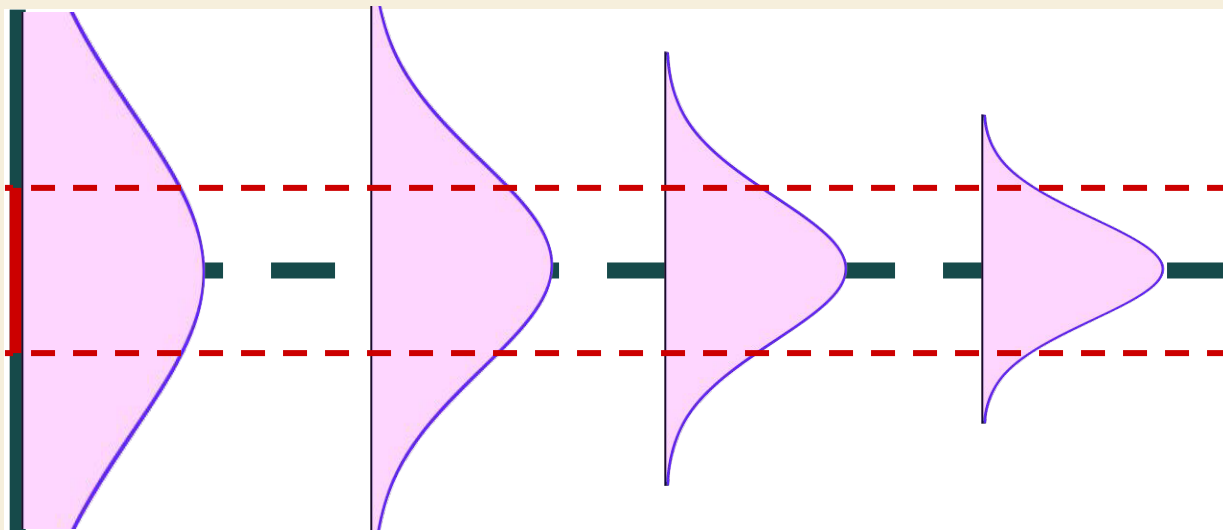


Convergence of sum of rv

Convergence of Sequence of R.V.:

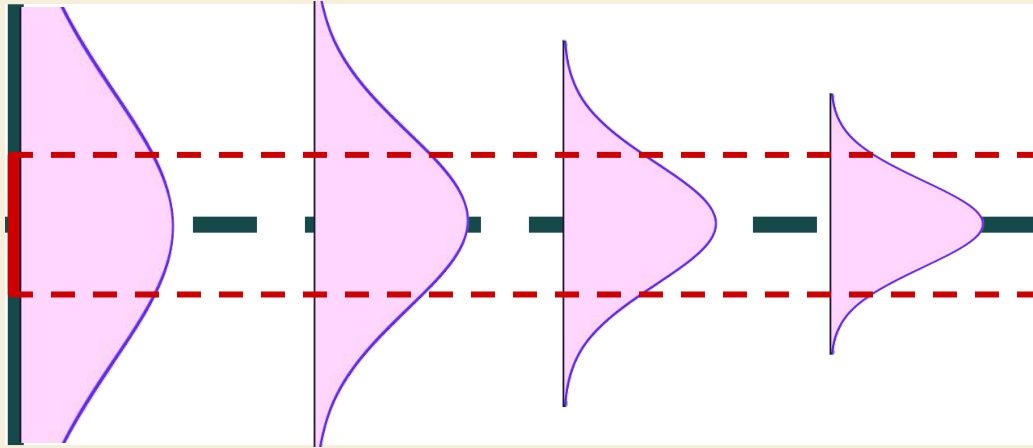
$$\forall \epsilon > 0, \lim_{n \rightarrow \infty} P(|Z_n - a| \geq \epsilon) = 0$$

want prob of falling outside
band $\rightarrow 0$



Convergence of Sequence of R.V.:

$$\forall \epsilon > 0, \lim_{n \rightarrow \infty} P(|Z_n - a| \geq \epsilon) = 0$$



$$\forall \epsilon > 0, \forall \epsilon' > 0, \exists M \text{ s.t. } \forall M' > M, P(|Z_{n_{M'}} - a| \geq \epsilon) \leq \epsilon'$$

↑
red band

↑
there exist
 n

↑
every distribution
after n

↑
probability falling
outside red band

Convergence of our error estimate

$$\hat{Z} = \frac{Z_1 + Z_2 + \dots + Z_N}{N}$$

Convergence of our error estimate

$$\hat{Z} = \frac{Z_1 + Z_2 + \dots + Z_N}{N}$$

$$\mathbb{E}[\hat{Z}] = \frac{\mathbb{E}[Z_1] + \mathbb{E}[Z_2] + \dots + \mathbb{E}[Z_N]}{N}$$

Convergence of our error estimate

$$\hat{Z} = \frac{Z_1 + Z_2 + \dots + Z_N}{N}$$

$$\mathbb{E}[\hat{Z}] = \frac{\mathbb{E}[Z_1] + \mathbb{E}[Z_2] + \dots + \mathbb{E}[Z_N]}{N}$$

$$= \frac{\mu_1 + \mu_2 + \dots + \mu_N}{N} = \mu \quad \text{as i.i.d. distributed}$$

Convergence of our error estimate

$$\hat{Z} = \frac{Z_1 + Z_2 + \dots + Z_N}{N}$$

$$\mathbb{E}[\hat{Z}] = \frac{\mathbb{E}[Z_1] + \mathbb{E}[Z_2] + \dots + \mathbb{E}[Z_N]}{N}$$

$$= \frac{\mu_1 + \mu_2 + \dots + \mu_N}{N} = \mu$$

$$\mathbb{V}[\hat{Z}] = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

$$\begin{aligned} V(\frac{1}{N}) &= V(\frac{Z_1}{N}) + V(\frac{Z_2}{N}) + \dots \\ &= \frac{1}{N^2} V(Z_1) + \frac{1}{N^2} V(Z_2) + \dots \\ &= \frac{1}{N^2} (\sigma^2 + \sigma^2 + \dots) \\ &= \frac{1}{N^2} \times N \sigma^2 \\ &= \frac{\sigma^2}{N} \end{aligned}$$

Convergence of our error estimate

$$\hat{Z} = \frac{Z_1 + Z_2 + \dots + Z_N}{N}$$

$$\mathbb{E}[\hat{Z}] = \frac{\mathbb{E}[Z_1] + \mathbb{E}[Z_2] + \dots + \mathbb{E}[Z_N]}{N}$$

$$= \frac{\mu_1 + \mu_2 + \dots + \mu_N}{N} = \mu$$

$$\mathbb{V}[\hat{Z}] = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{\boxed{N}}$$

Looks like our
variance is getting
smaller!

Plugging in our results to Chebyshev

$$\frac{\sigma^2}{a^2} \geq P(|z - \mu| \geq a)$$

Note this isn't exactly Chebyshev, but it the step just before (so equivalent)

$$\frac{1}{k^2} \geq P(|z - \mu| \geq k\sigma), \quad a = k\sigma \rightarrow k = \frac{a}{\sigma}$$

Plugging in our results to Chebyshev

$$\frac{\sigma^2}{a^2} \geq P(|z - \mu| \geq a)$$

Note this isn't exactly Chebyshev, but it the step just before (so equivalent)

$a \rightarrow \epsilon, \sigma^2 \rightarrow V(\hat{Z})$:

$$P(|\hat{Z} - \mu| \geq \epsilon) \leq \frac{\mathbb{V}[\hat{Z}]}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2}$$

We call this the weak law of large numbers

$$\frac{\sigma^2}{a^2} \geq P(|z - \mu| \geq a)$$

Note this isn't exactly Chebyshev, but it the step just before (so equivalent)

$$P(|\hat{Z} - \mu| \geq \epsilon) \leq \frac{\mathbb{V}[\hat{Z}]}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2}$$

Weak law of large numbers

$$P(|\hat{Z} - \mu| \geq \epsilon) \leq \frac{\mathbb{V}[\hat{Z}]}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2}$$

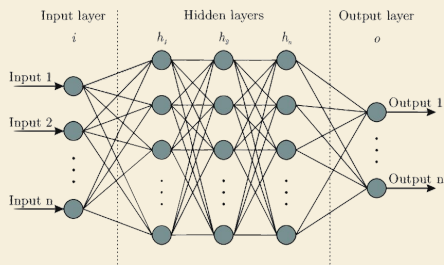
As N gets bigger, our sample error estimate gets closer to the true mean (converge in the probability sense we gave), so we do not need to compute the integral to get our expectation, we can use the sample mean (this is just Monte Carlo integration)

$$P(|\hat{z} - m| > \epsilon) < \delta$$

↑
error
↑
confidence

A final word: No free lunch!

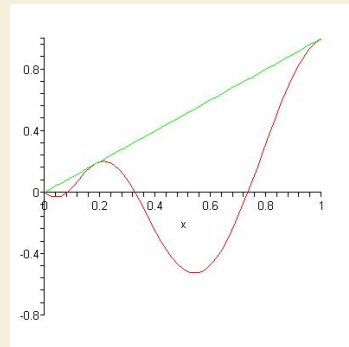
In machine learning, if you have a uniform distribution over off-training set examples, no one classifier will have an a priori advantage in its generalization error!



Definition 2.2. Universal Approximation Theorem (Cybenko, 1989) Let $\sigma(x)$ be a sigmoidal activation function, i.e., a function that is nonconstant, bounded, and continuously differentiable. For any continuous function f on a compact subset of \mathbb{R}^n , and any $\varepsilon > 0$, there exist positive integers N , weights w_i , and biases b_i , and a sum of N sigmoidal functions such that:

$$F(x) = \sum_{i=1}^N w_i \sigma(w_i^T x + b_i)$$

satisfies $|F(x) - f(x)| < \varepsilon$ for all x in the compact subset.



Definition 2.1. Stone-Weierstrass Theorem Let f be a continuous function on a closed interval $[a, b]$. For any $\varepsilon > 0$, there exists a polynomial function $P(x)$ such that

$$\|f - P\|_{\infty} = \sup_{x \in [a, b]} |f(x) - P(x)| < \varepsilon.$$

In other words, the polynomial $P(x)$ can uniformly approximate f on the interval $[a, b]$.



Next lecture: Advanced Concentration and Cross-Validation

