A Mathematical Perspective on Machine Learning

Weinan E

Center for Machine Learning Research and
School of Mathematical Sciences

Peking University

(z)activation function

goal

average error for all the data

m:number of free parameters

square error

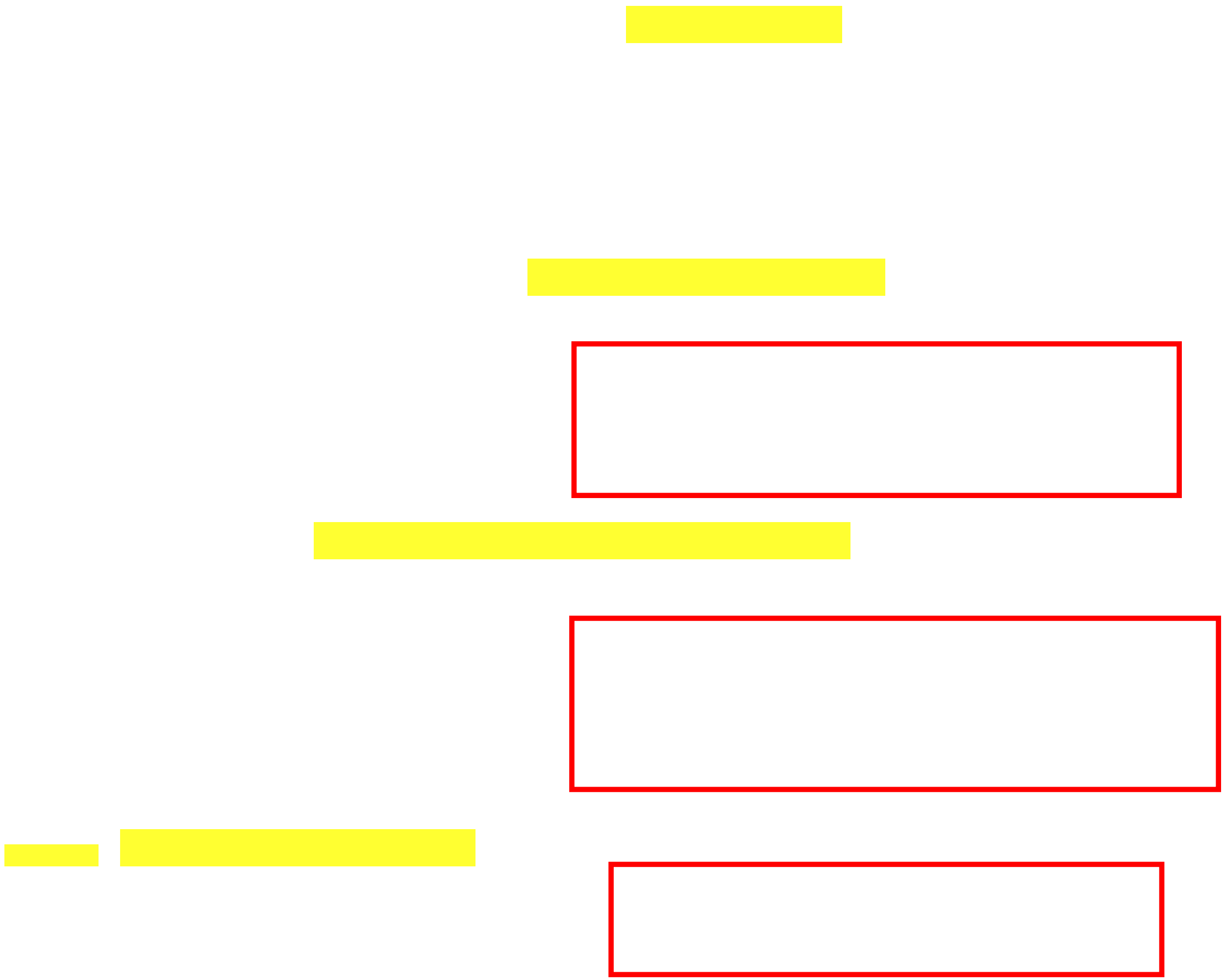
when n is large, use a ran

is a property distribution, so can be written as an expectation

approximate, finite sum

doesn't work in high dimention

convergence rate is independent of dimention



"New" approach: Let π be a probability distribution

repalce dω   written as an expectation

$$f^*(\boldsymbol{x}) = \int_{\mathbb{R}^d} a(\boldsymbol{\omega}) e^{i(\boldsymbol{\omega},\boldsymbol{x})} \pi(d\boldsymbol{\omega}) = \boxed{\mathbb{E}_{\boldsymbol{\omega}\sim\pi} a(\boldsymbol{\omega}) e^{i(\boldsymbol{\omega},\boldsymbol{x})}}$$

Let $\{\boldsymbol{\omega}_j\}$ be an **i.i.d. sample of π**, $\boxed{f_m(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^m a(\boldsymbol{\omega}_j) e^{i(\boldsymbol{\omega}_j,\boldsymbol{x})}}$,

$$\boxed{\mathbb{E}|f^*(\boldsymbol{x}) - f_m(\boldsymbol{x})|^2 = \frac{\text{var}(f)}{m}}$$ does not suffer from CoD

Note: $f_m(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^m a_j \sigma(\boldsymbol{\omega}_j^T \boldsymbol{x})$ = two-layer neural network with $\sigma(z) = e^{iz}$.

Conclusion:

Functions of the this type (i.e. can be expressed as this kind of expectation) can be approximated by two-layer neural networks with a dimension-independent error rate.

Approximation theory for the *random feature model*

ω：feature vector
Ω：sapce of feature vector

- Let $\phi(\cdot;\boldsymbol{w})$ be a feature function parametrized by $\boldsymbol{w} \in \Omega$, e.g. $\phi(\boldsymbol{x};\boldsymbol{w}) = \sigma(\boldsymbol{w}^T\boldsymbol{x})$.
  We will assume that ϕ is continuous and Ω is compact.
- Let π_0 be a fixed distribution for the random variable $\boldsymbol{w}$.
- Let $\{\boldsymbol{w}_j^0\}_{j=1}^m$ be a set of i.i.d samples drawn from π_0.

The random feature model (RFM) associated with the features $\{\phi(\cdot;\boldsymbol{w}_j^0)\}$ is given by

$$\boxed{f_m(\boldsymbol{x};\boldsymbol{a}) = \frac{1}{m}\sum_{j=1}^{m} a_j\phi(\boldsymbol{x};\boldsymbol{w}_j^0)}.$$

linear combination with aj

What spaces of functions are "well approximated" (say with the same convergence rate as in Monte Carlo) **by the random feature model?**

- In classical approximation theory, these are a the Sobolev or Besov spaces: They are characterized by the convergence behavior for some specific approximation schemes.
- Direct and inverse approximation theorems.

Define the kernel function associated with the random feature model:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w} \sim \pi_0}[\phi(\boldsymbol{x}; \boldsymbol{w})\phi(\boldsymbol{x}'; \boldsymbol{w})]$$

Let $\mathcal{H}_k$ be the reproducing kernel Hilbert space (RKHS) induced by the kernel k.

Probabilistic characterization: a(.): coefficient function

$f \in \mathcal{H}_k$ if and only if there exists $a(\cdot) \in L^2(\pi_0)$ such that

$$\boxed{f(\boldsymbol{x}) = \int a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})d\pi_0(\boldsymbol{w})} = \boxed{\mathbb{E}_{\boldsymbol{w} \sim \pi_0} a(\boldsymbol{w})\phi(\boldsymbol{x}; \boldsymbol{w})}$$

and

$$\boxed{\|f\|^2_{\mathcal{H}_k}} = \boxed{\int a^2(\boldsymbol{w})d\pi_0(\boldsymbol{w})} = \boxed{\mathbb{E}_{\boldsymbol{w} \sim \pi_0} a^2(\boldsymbol{w})}$$

sqaure of l2 norm

the rest in : 2022_ICM_A Mathematical Perspective of Machine Learning_note2