Applications of Fourier analytic Barron Space theory

Motivation

Jonathan Miller, April 24, 2024

Applications of Fourier analytic Barron Space Theory

Motivation: Outline

- Introduction
  - Generalization Error
- Literature
- Barron space
- Approximation Error
- Estimation Error
- Fourier analytic Barron Space Theory

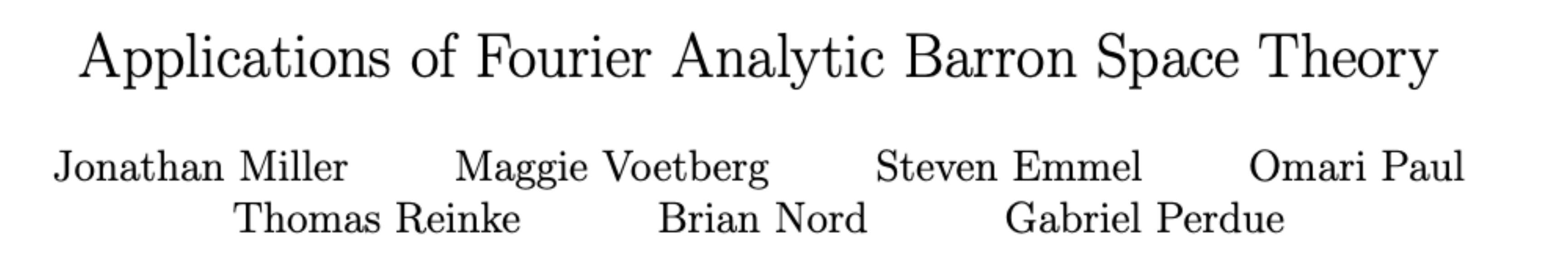
Introduction

Euler Scientific is a small research company.

Collaboration with Fermilab/NGA on Bounding Generalization Risk for Deep Neural Networks.

Today I will talk about applications for shallow Neural Networks that came out of that collaboration.

The is about Euler Scientific's gbtoolbox and not on behalf of the collaboration (publications forthcoming).

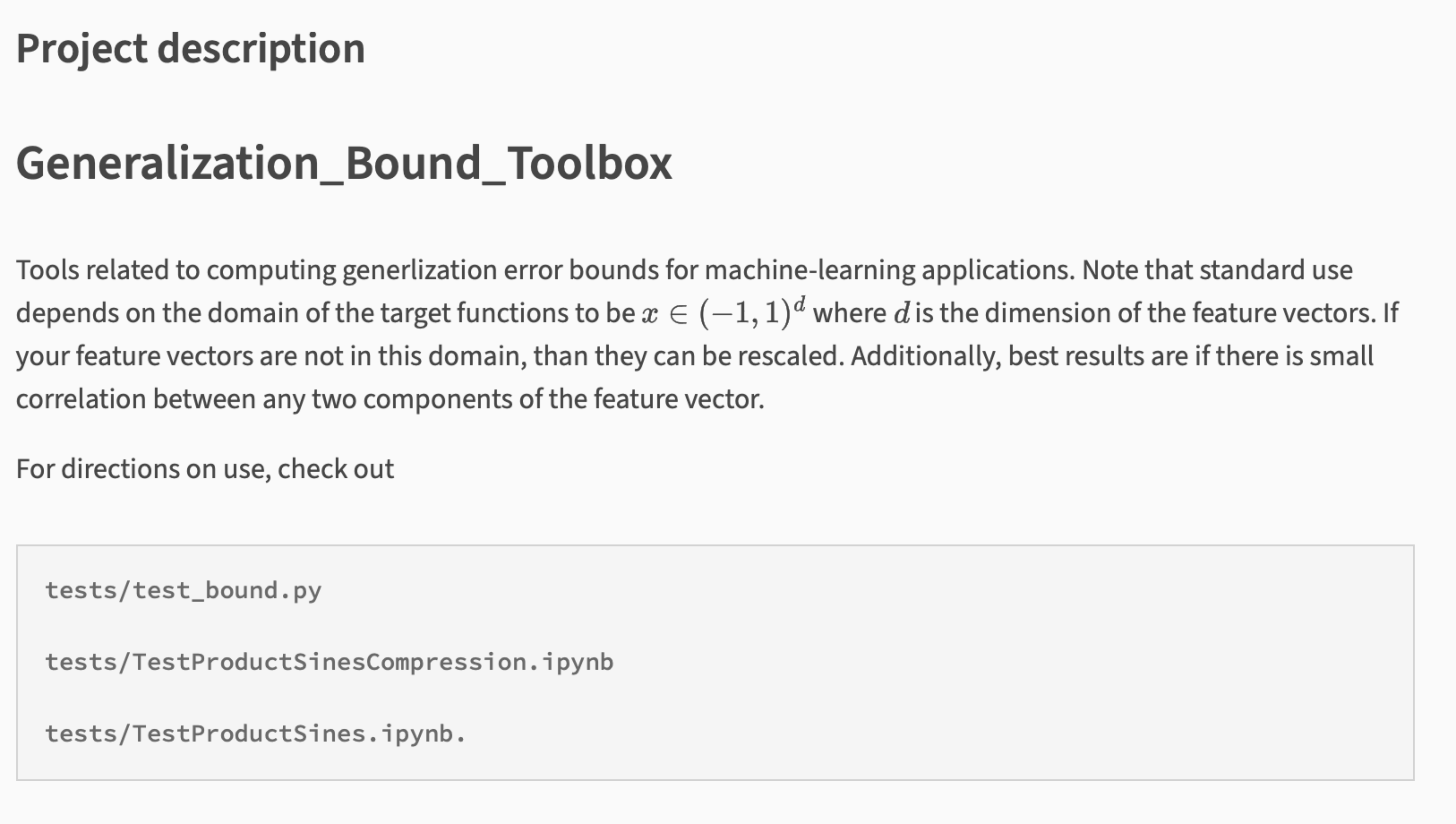


Applications of Fourier Analytic Barron Space Theory

Jonathan Miller   Maggie Voetberg   Steven Emmel   Omari Paul
Thomas Reinke   Brian Nord   Gabriel Perdue

Applications of Fourier analytic Barron Space Theory

Generalization Bound Toolbox (on PyPI)



Project description

Generalization_Bound_Toolbox

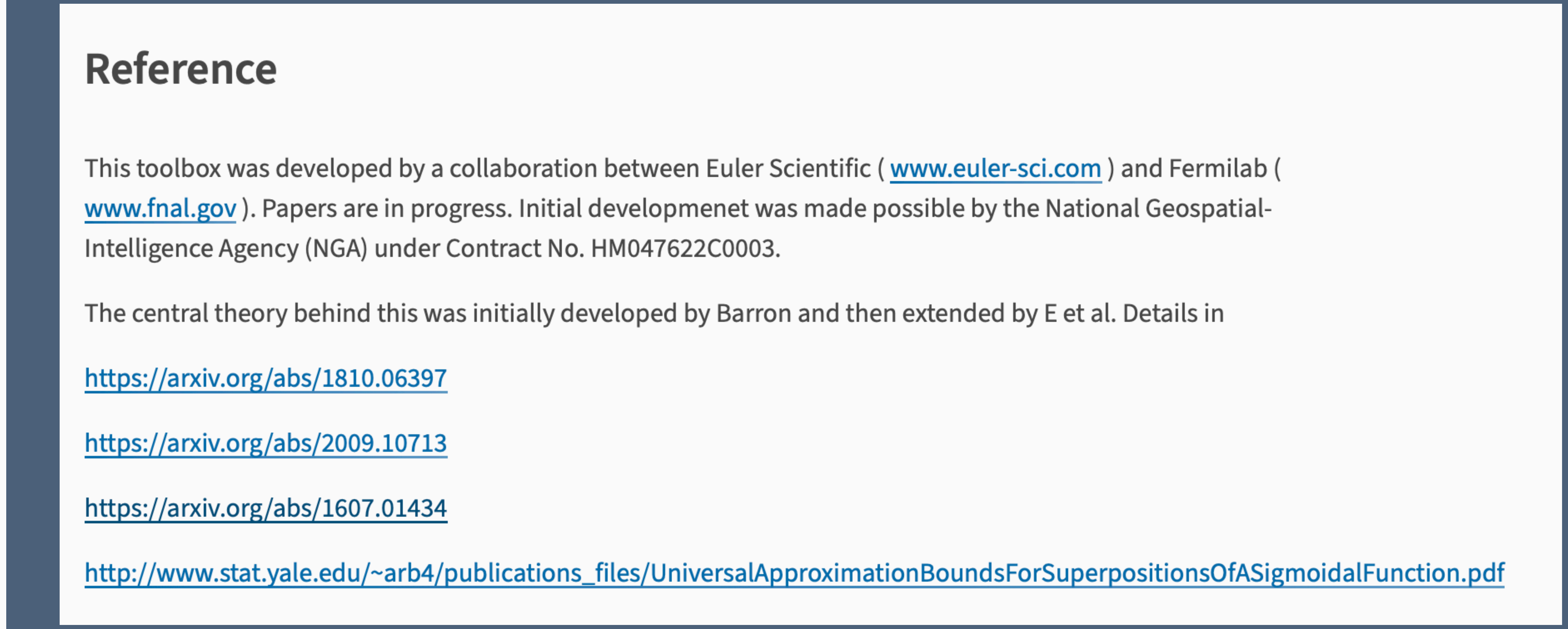
Tools related to computing generlization error bounds for machine-learning applications. Note that standard use depends on the domain of the target functions to be $x \in (-1,1)^d$ where d is the dimension of the feature vectors. If your feature vectors are not in this domain, than they can be rescaled. Additionally, best results are if there is small correlation between any two components of the feature vector.

For directions on use, check out

```
tests/test_bound.py

tests/TestProductSinesCompression.ipynb

tests/TestProductSines.ipynb.
```



Reference

This toolbox was developed by a collaboration between Euler Scientific (www.euler-sci.com) and Fermilab (www.fnal.gov). Papers are in progress. Initial developmenet was made possible by the National Geospatial-Intelligence Agency (NGA) under Contract No. HM047622C0003.

The central theory behind this was initially developed by Barron and then extended by E et al. Details in

https://arxiv.org/abs/1810.06397

https://arxiv.org/abs/2009.10713

https://arxiv.org/abs/1607.01434

http://www.stat.yale.edu/~arb4/publications_files/UniversalApproximationBoundsForSuperpositionsOfASigmoidalFunction.pdf

https://pypi.org/project/gbtoolbox/

Fourier analytic Barron Space theory

Introduction

- We have a complex phenomena.
- We postulate some function, that takes the data we have and provides an output.
- We find some approximate function.
- We apply it to some new data, that we assume to be the same complex phenomena (so shares the same function).
- How well does the approximate function describe the new data?

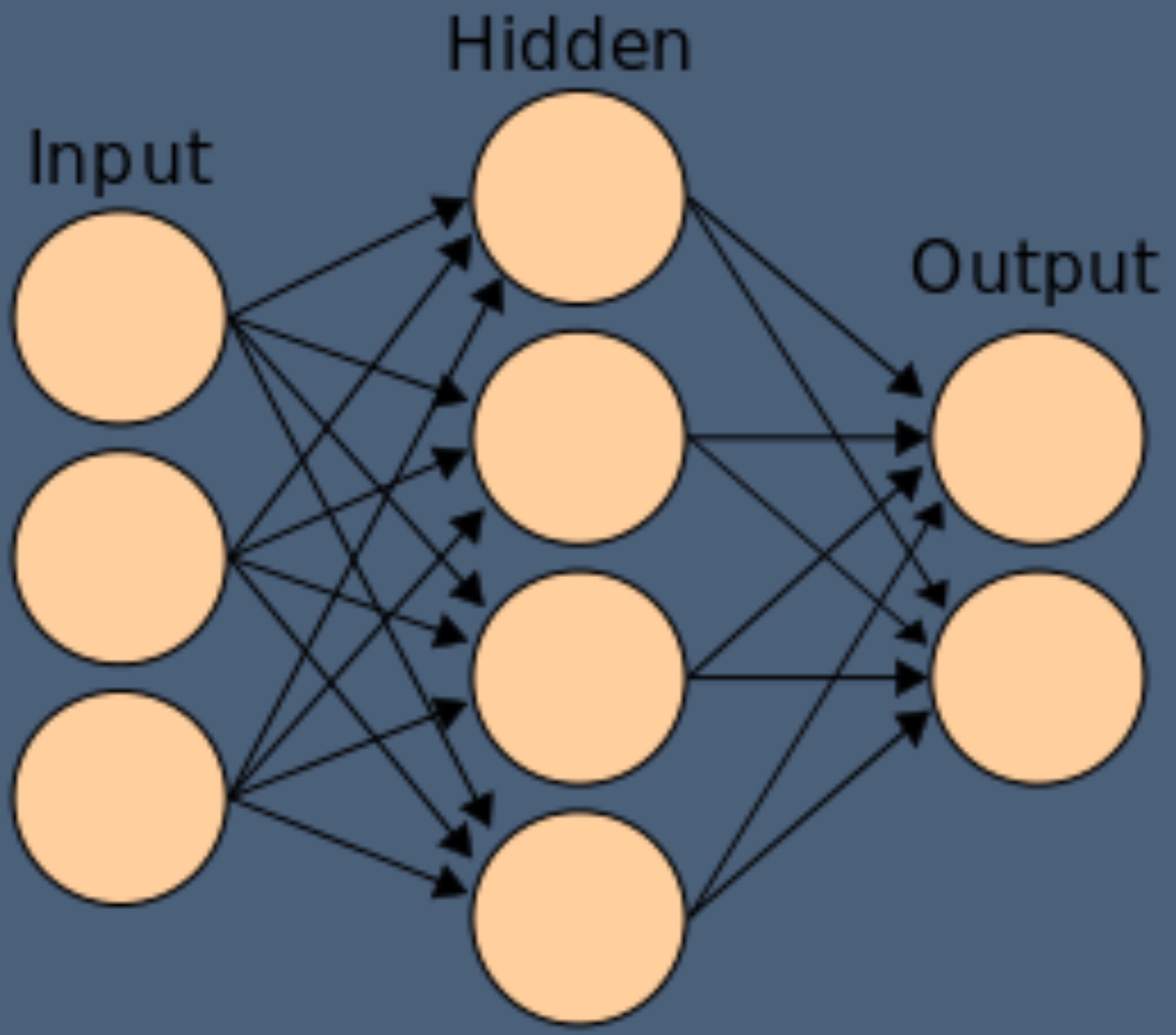


Image from Wikipedia

Fourier analytic Barron Space theory

Generalization Error

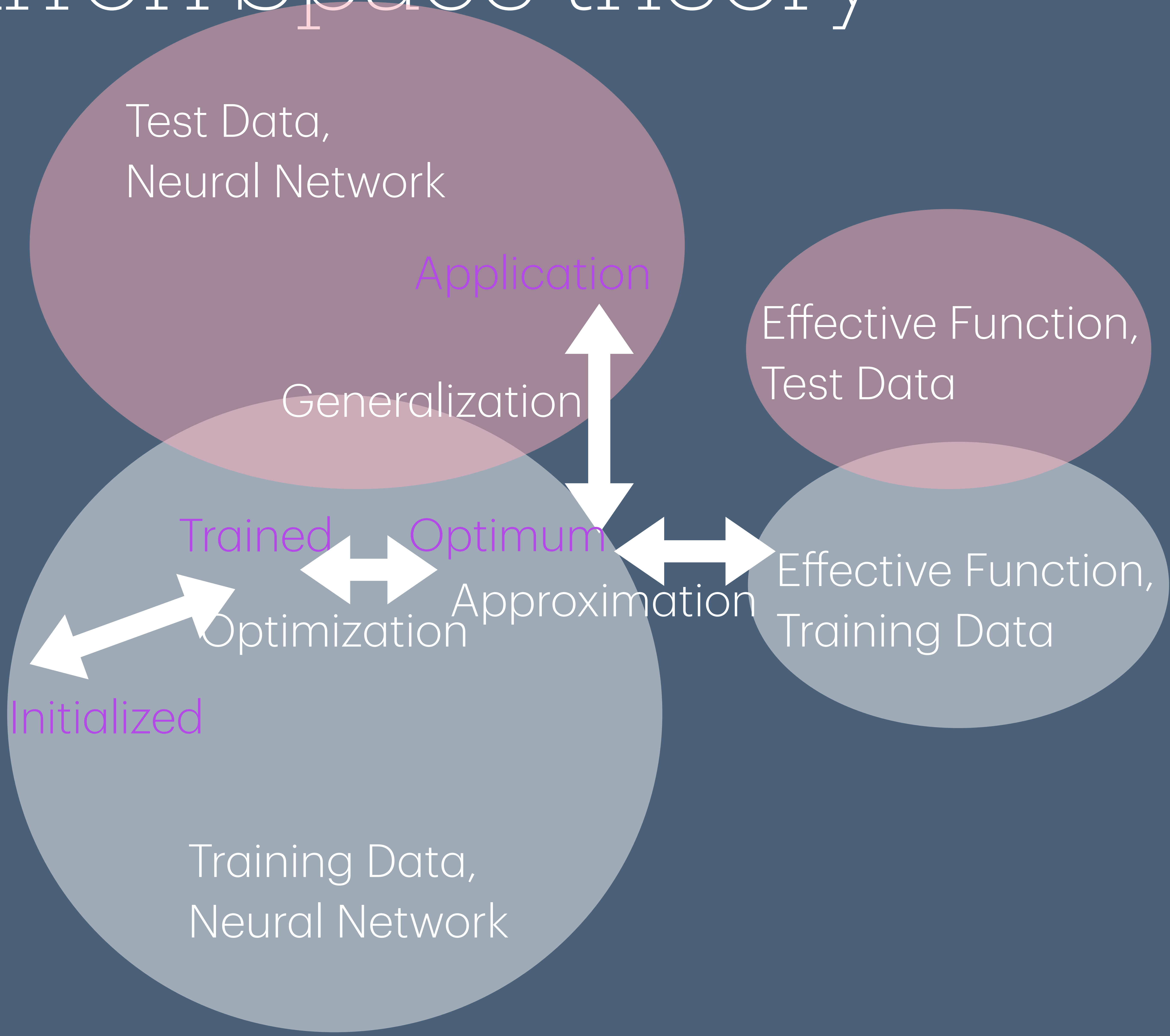
- We are concerned with the error associated with the difference between trained network on the training data and on unseen data in the same domain.

- We can divide this error into the difference between the effective target function (which depends on the training data) and the function space defined by the neural network, or Approximation Error.

  - And Estimation Error (Sometimes also called Generalization Error), or the difference due to picking some other examples.

  - And Optimization Error, or finding the best Approximation from some Initialized Network.

- Generalization Error = Approximation Error + Estimation Error + Optimization Error



Fourier analytic Barron Space theory

Literature Review

- A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," in IEEE Transactions on Information Theory, vol. 39, no. 3, pp. 930-945, May 1993

- L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," in IEEE Transactions on Information Theory, vol. 39, no. 3, pp. 999-1013, May 1993

- Weinan E et al., “A priori estimates of the population risk for two-layer neural networks,” in Communications in Mathematical Sciences, vol. 17, no. 5, pp. 1407-1425, 2019

- A. R. Barron and J. M. Klusowski, “Approximation and Estimation for High-Dimensional Deep Learning Networks,” 1809.03090 (unpublished)

- J. M. Klusowski and A. R. Barron, “Risk Bounds for High-dimensional Ridge Function Combinations Including Neural Networks,” 1607.01434 (unpublished)

- Weinan E et al., “Towards a Mathematical Understanding of Neural Network-Based Machine Learning: What We Know and What We Don’t,” in CSIAM Transactions on Applied Mathematics, vol. 1, no. 4, pp. 561-615, 2020

- S. Shalev-Shwartz and S. Ben-David, "Understanding Machine Learning: From Theory to Algorithms,” USA: Cambridge University Press, 2014

- Behnam Neyshabur et al., “Norm-Based Capacity Control in Neural Networks,” in Proceedings of The 28th Conference on Learning Theory, PMLR 40:1376-1401, 2015

Fourier analytic Barron Space theory

Barron Space

- We consider functions $f(\mathbf{x})$, $\mathbf{x} \in [-1,1]^d$ and we consider an effective target function, $f^*(\mathbf{x})$
- $f^*(\mathbf{x})$ is the Fourier transform of $\tilde{f}^*(\omega)$, we also define $\|\omega\|_1 = \sum_j^d |\omega_j|$
  - Later we will talk about approximate of $\tilde{f}^*(\omega)$.
- Then we can define the Barron norm $\gamma(f^*) = \inf_{f^*} \int \|\omega\|_1^2 |\hat{f}^*(\omega)| \, d\omega < \infty$
  - This defines our the function space that we are interested in, the paradigm that Fourier analytic Barron space theory works.
  - Recall also the manifold hypothesis, while $\gamma(f^*)$ increase with d in practice it isn't going to be the maximum.

Fourier analytic Barron Space theory

Barron Space

$$\gamma(f^*) = \inf_{f^*} \int \|\omega\|_1^2 |\hat{f}^*(\omega)| \, d\omega < \infty$$

- Problem: Isn't the Barron Norm very large? A lot of functions we want to approximate we imagine to have discontinuities.

- However, we don't have an analytic function, we have some set of $\{\mathbf{x}_i, y_i\}$. Our function space that we are finding the minimum $\gamma(f^*)$ over are those such that $f^*(\mathbf{x}_i) = y_i$.
  - Note this is also the case for Test, $f^*(\mathbf{x}_j) = y_j$.

- So given some finite discontinuities (or even not enough smoothness), there exists some $f^*(\mathbf{x}_i)$ such that there is a $\gamma(f^*) < \infty$.
  - Differences are going to be Estimation Error (or Error due to different Domains).

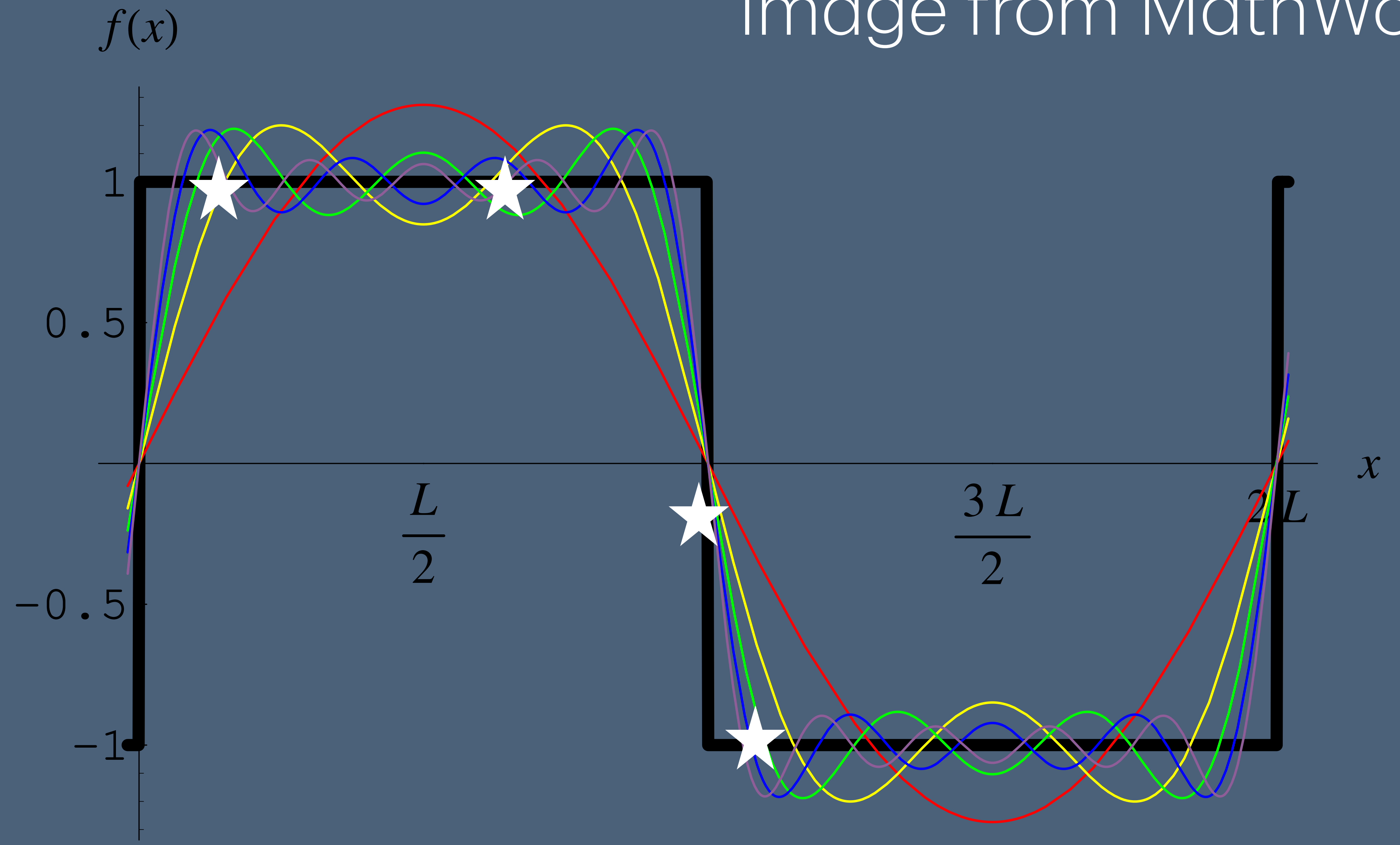


Image from MathWorld

Fourier analytic Barron Space theory

Barron Space

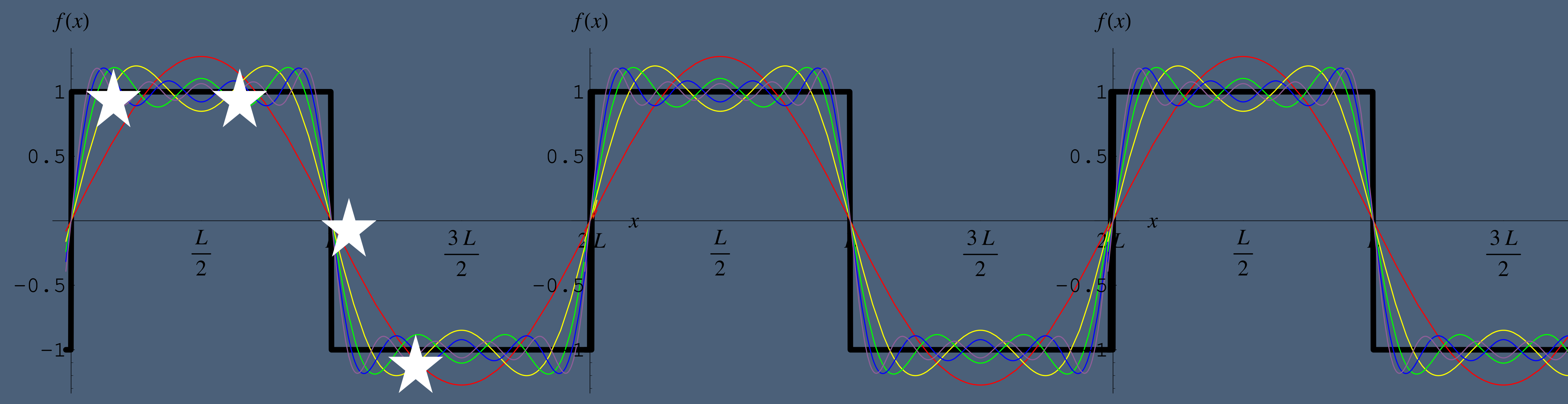
$$\gamma(f^*) = \inf_{f^*} \int \|\omega\|_1^2 |\hat{f}^*(\omega)| \, d\omega < \infty$$

- Problem: how to define a Fourier transform on $\mathbf{x} \in [-1,1]^d$?

- Well, we are extending to $\mathbf{x} \in [-\infty, \infty]^d$. The obvious extension for analytic, isn't going to be the extension that minimizes the Barron Norm.

- Our effective target function, the one that we are approximating with our neural network, is the extension that minimizes the Barron Norm.

Image from MathWorld



Fourier analytic Barron Space theory

Motivation Approximation

- $f(\mathbf{x}) \simeq \int\int e^{i\omega\mathbf{x}}e^{-i\omega\mathbf{y}}f(\mathbf{y})d\omega d\mathbf{y}$ this is just the Fourier transform of a Fourier transform.
- $f(\mathbf{x}) \simeq \int e^{i\omega\mathbf{x}}\tilde{f}(\omega)d\omega$ then we do the Fourier transform (recall $\|\omega\|_1 = \sum_j^d |\omega_j|$)
    - Now we recall the Fourier transform of $\int \sigma(\hat{\omega}\mathbf{x}+t)e^{i\|\omega\|_1 t}dt \simeq \frac{e^{i\omega\mathbf{x}}}{\|w\|_1^2} + \delta(\omega)$
- $f(\mathbf{x}) \simeq \int\int \sigma(\hat{\omega}\mathbf{x}+t)e^{i\|\omega\|_1 t}\tilde{f}(\omega)\|\omega\|_1^2 d\omega dt$ assuming that $f(0) = 0$ and $\nabla f(0) = 0$ (note that $\tilde{f}(\omega) = |\tilde{f}(\omega)|e^{i\mathrm{Arg}(f)}$)
- Consider MC estimator $f(\mathbf{x}) \simeq \sum_i^M \frac{\rho(\omega_i, t_i)}{M}\sigma(\hat{\omega}_i\mathbf{x}+t_i)$ where come from PDF $|\tilde{f}(\omega)||\|\omega\|_1^2|\cos(\|\omega\|t+\mathrm{Arg}(f))|$

Fourier analytic Barron Space theory

Approximation Error: A bit more formal

- We have a neural network $\sum_{i}^{M} a_i\sigma(b_i\mathbf{x} + c_i)$ and can define the Path Norm $\|\Theta\|_P = \sum_{k}^{m} |a_k|(\|\mathbf{b}_k\|_1 + |c_k|)$

- We consider the MC estimator, as motivated, $\hat{f}_m(\mathbf{x}, \{\omega, \beta, z\}) = \frac{1}{m}\sum_{j}^{m} \rho(\omega_j, \beta_j, z_j)\sigma(\hat{\omega}_j \cdot \mathbf{x} + z_j\beta_j)$ where

  $\{\omega_j, \beta_j, z_j\}$ come from the PDF $p(\omega, \beta, z) = |\hat{f}^*(\omega)|\|\omega\|_1^2|\cos(\|\omega\|_1\beta z - \arg(\hat{f}^*(\omega))|/\nu$ where $\nu \leq 2\gamma(f^*)$.

  - We have a bound on the error of an MC estimator.

- E et al. proved that there are paramete such that both $\|\Theta\|_P \leq 2\gamma_2(f^*)$ and

  $\mathrm{E}_x[(f(x) - f_m(x, \{\omega, t, z\}))^2] \leq \frac{3\gamma_2^2(f^*)}{m}$

Fourier analytic Barron Space theory

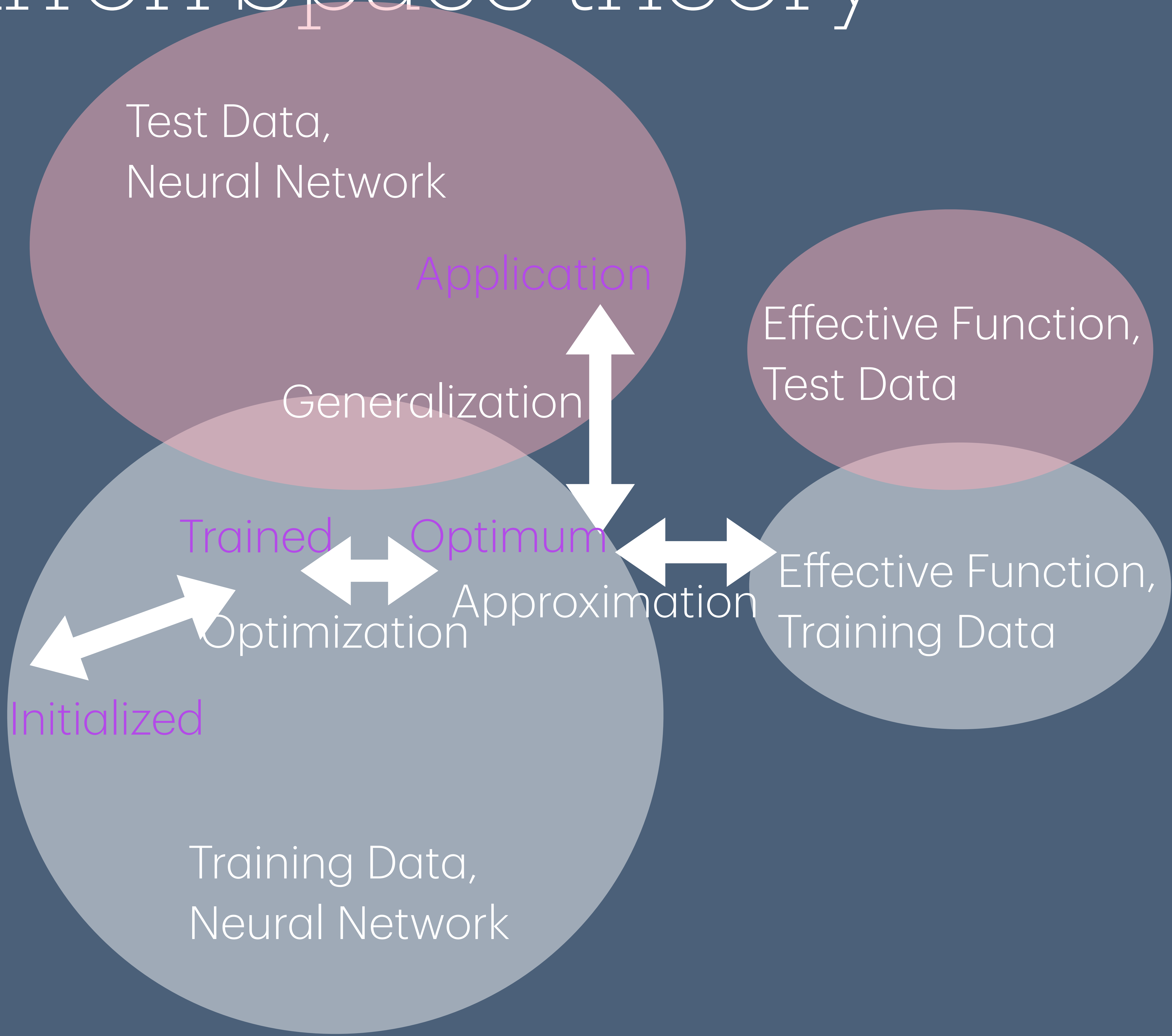
Estimation Error

- Consider some regularized neural network (with L1 regularization λ, but both I and E et al. find this not important).
- If we consider all the f_m that have $\|\Theta\|_P \leq Q$, these f_m are a hypothesis space F. The Rademacher complexity of this space, $R_n(Q) \leq 2Q\sqrt{\frac{2\ln(2d)}{n}}$. And $R_n = \frac{1}{n}\mathrm{E}_\epsilon[\sup_{f\in F}\sum_i^n \epsilon_i f(x_i)]$ where $\{\epsilon_i\}$ are $\pm 1/2$.
- For $\delta > 0$, with probability $1-\delta$ over the choice of $Z = \{z_k\}$, $|\frac{1}{n}\sum_i^n f(x_i) - \mathrm{E}_x[f(x)]| \leq \mathrm{E}_Z[R_n(Q)] + B\sqrt{\frac{2\ln(2/\delta)}{n}}$ where $|f(z)| < B$
- Then we have $|L(\theta) - \hat{L}_n(\theta)| \leq 4A(\|\Theta\|_P + 1)\sqrt{\frac{2\ln(2d)}{n}} + B\sqrt{\frac{\ln(\sum_k 2k^{-2}(\|\Theta\|_P + 1)/\delta)}{n}}$ where $L(\theta)$ is the population risk and $\hat{L}_n(\theta)$ is the empirical risk.

Fourier analytic Barron Space theory

Generalization Error Bound

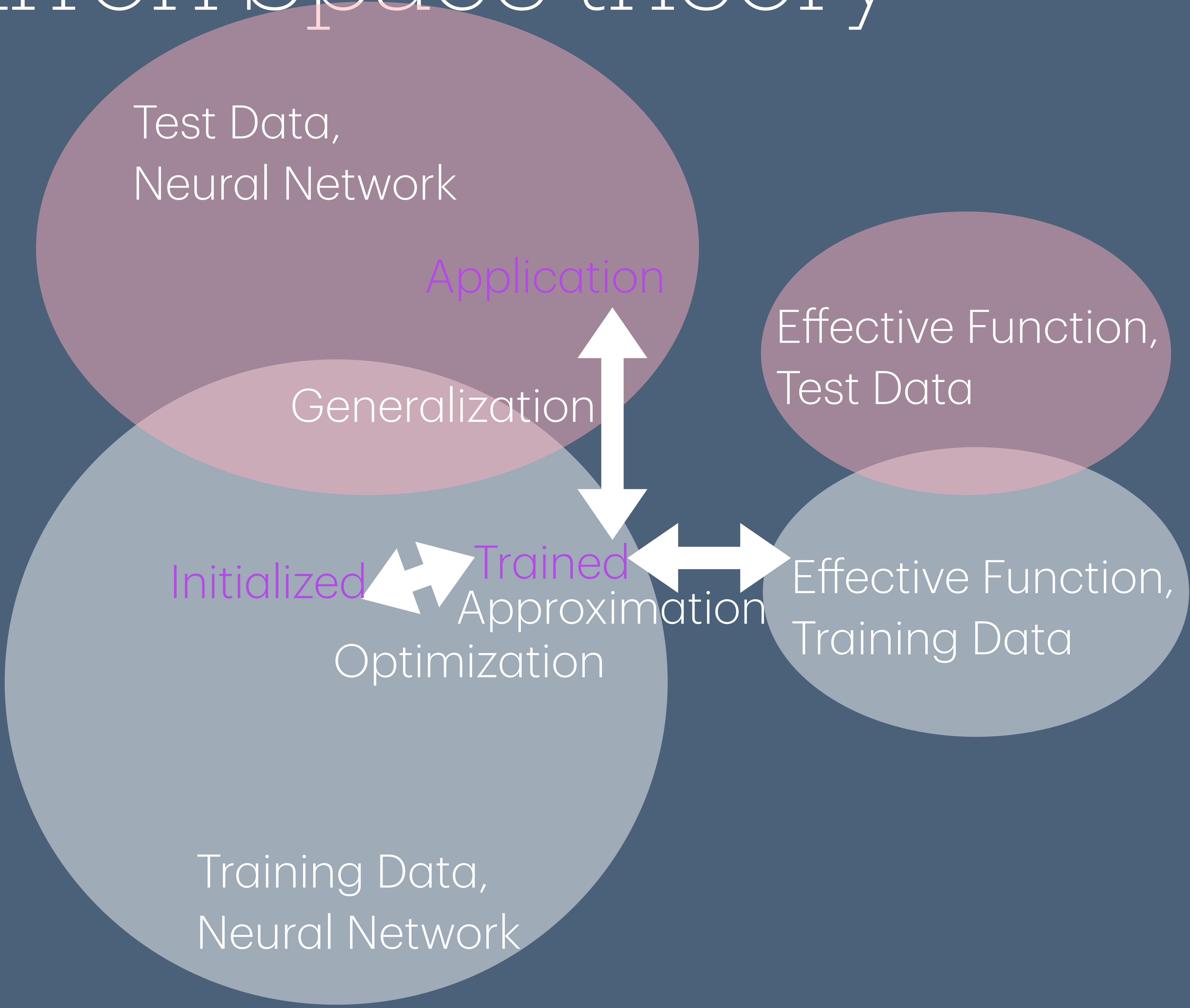
- We have Generalization Error = Approximation Error + Estimation Error.
  - Where the two Errors are connected through the Path Norm/Barron Norm.
- $E_x|f_m(\mathbf{x},\hat{\theta}) - f^*(x)|^2 \leq \frac{\gamma^2(f^*)}{m} + \lambda\tilde{\gamma}(f^*) + \left(\tilde{\gamma}(f^*) + \sqrt{\ln(n/\delta)}\right)/\sqrt{n}$ where $\tilde{\gamma}(f) = \max(\gamma(f),1)$, $\lambda \geq 4\sqrt{2\ln(2d)/n}$ (and is related to the effective regularization), $f^*(x)$ is the effective target function, d is the dimension, $\gamma(f)$ us the Barron norm and $1-\delta$ is the probability the bound holds.
- Thus we have a bound on the application of the optimum neural network on unseen in domain data.



Fourier analytic Barron Space theory

Discussion Generalization Error

- If we initialize close to the optimal, using this Barron Theory, the training is more likely to make us optimal.
  - If it doesn’t, the Error is at most can be determined from the initialized + training.
  - So we get an Error Bound, we possibly have a more precise network (from additional training).
  - We have some short training period, and possibly more successful, since we close.



Fourier analytic Barron Space theory

Discussion

- Memorization is approximation close to example data, generalization is approximation far from example data.

- Neural networks are statistical learners, in $\mathbf{X}$ but the approximation is also in $\boldsymbol{\omega}$ in the Fourier analytic Barron Space regime, which explains their success at generalization.

- We can determine how the distribution that reflects a neural network that approximates the effective target function changes due to removing/including data.

- It is clear that Fourier analytic Barron Space theory isn't the final theory of Neural Networks, but a step.

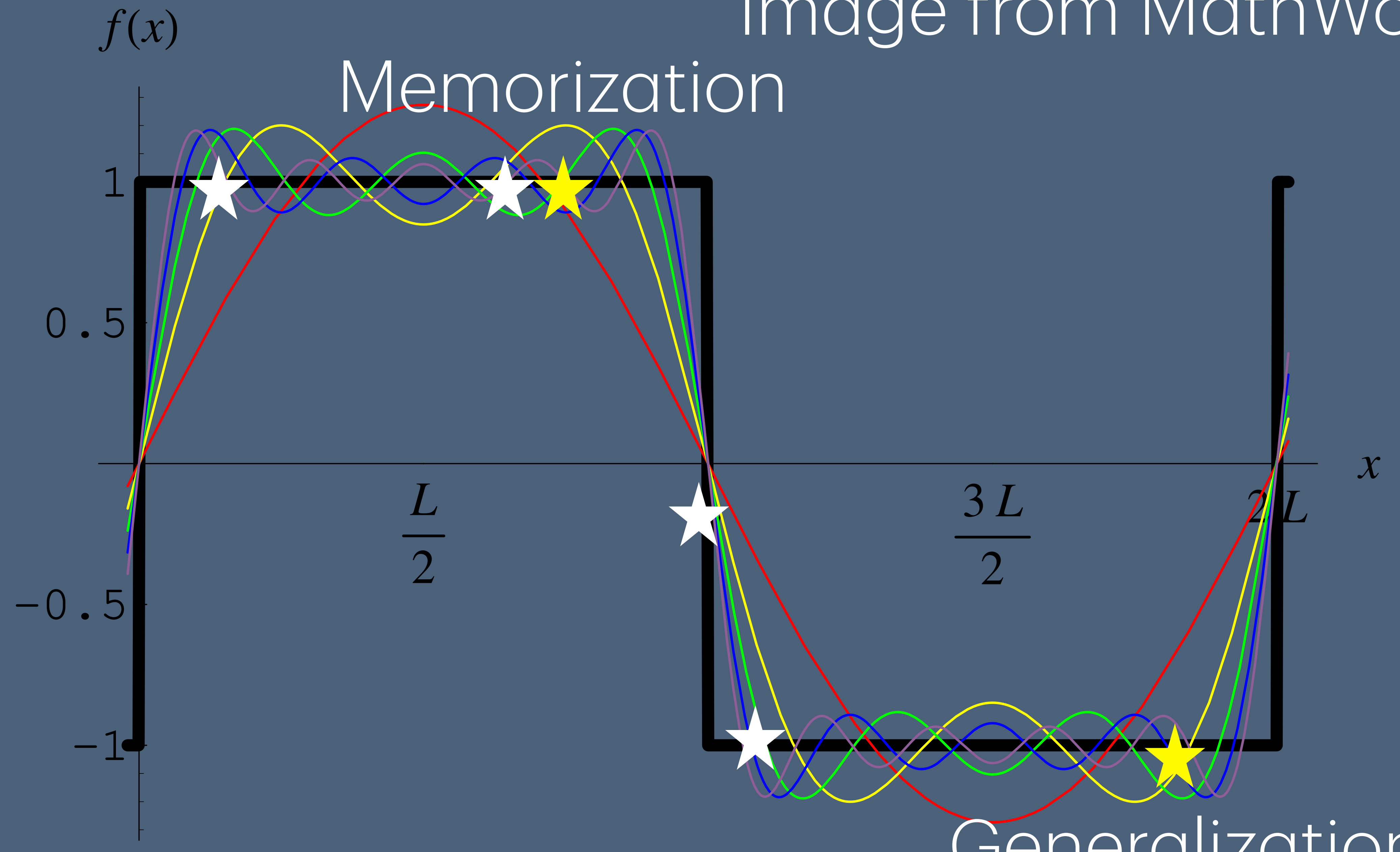


Image from MathWorld

Memorization

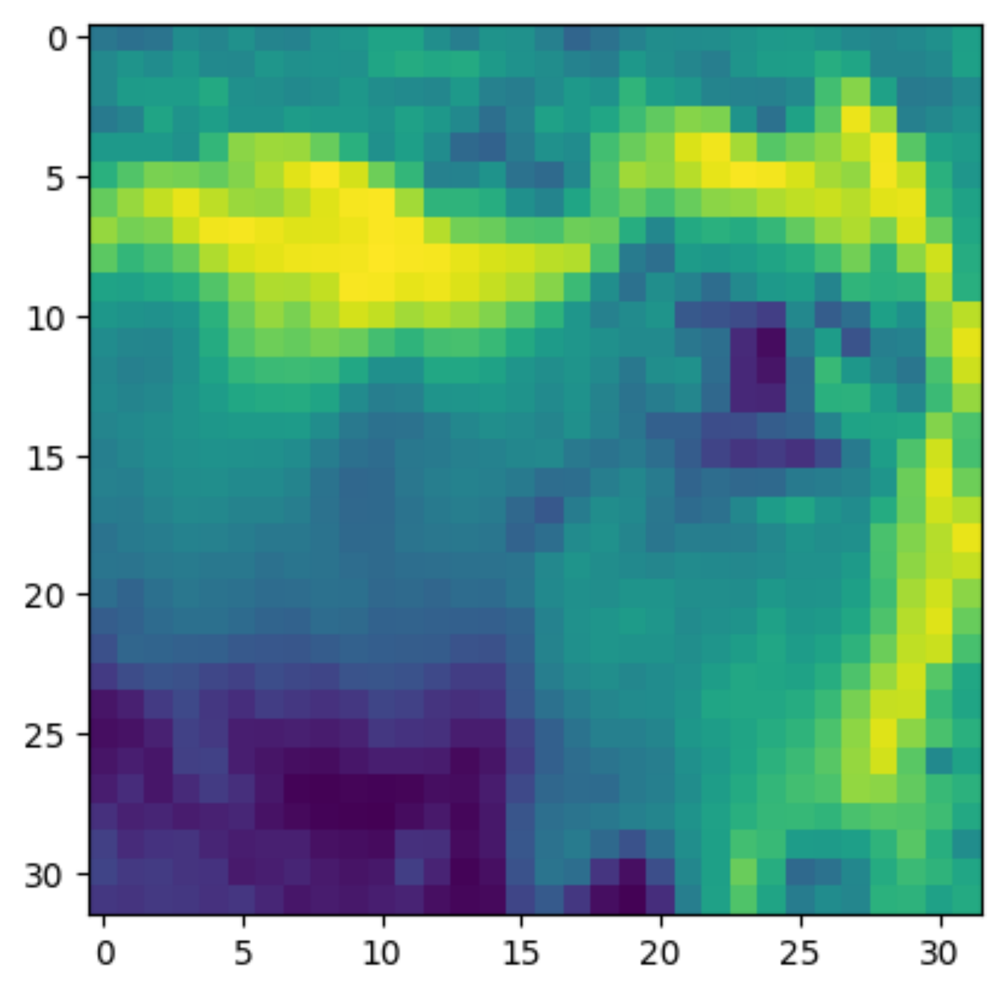
Generalization

Fourier analytic Barron Space theory

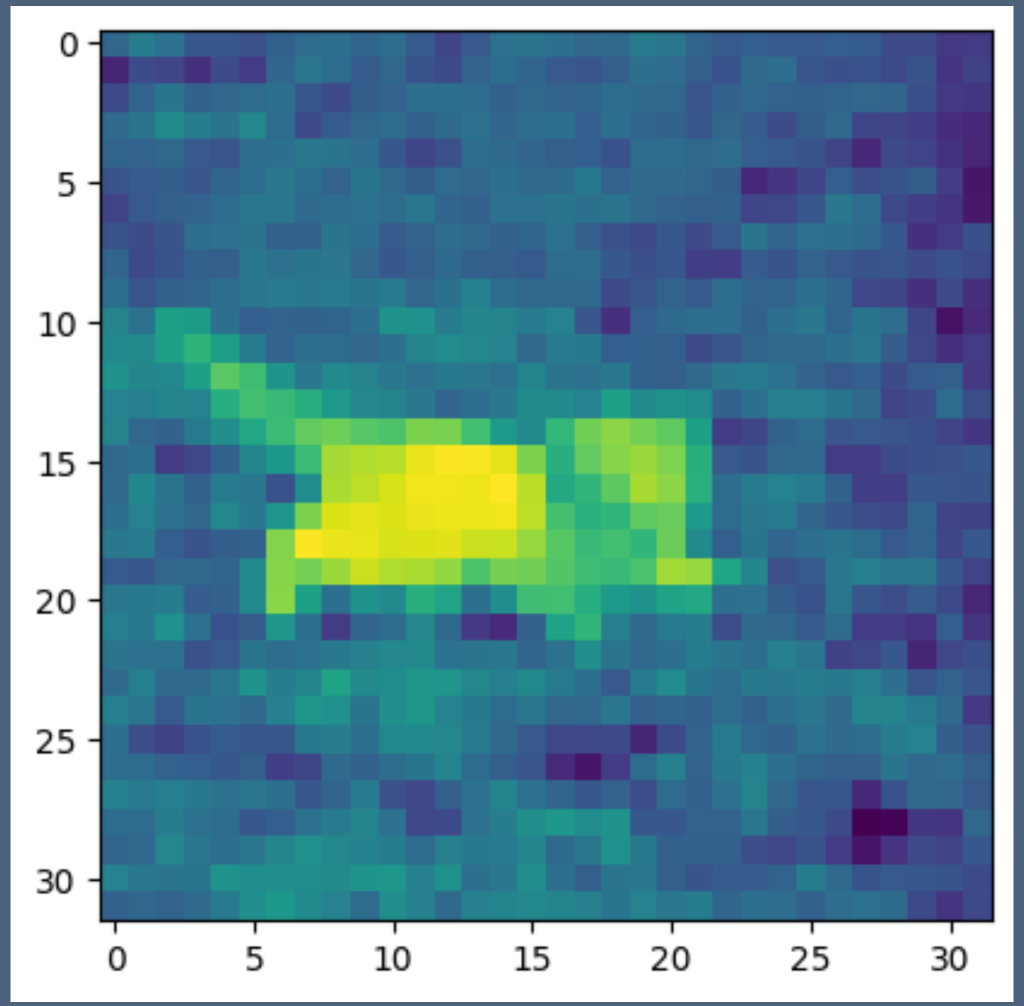
Discussion of Applications

- While the theory is a bit different for classification compared to regression, it is better to think of 'dogness' for example instead of 'is it a dog'.
- This is because the space that cifar10 exists in is $\left(2^{24}\right)^{1024}$ and even if we consider grey-scale cifar10 it is 2^{8192} . Even MNIST is in $2^{8\times28\times28=6272}$.
- Many points in this extended space might be dog-like but would have some 'dogness'.
- Probably some cases where the target function is fundamentally classification (0 or 1), but most cases we want our effective target function to be regression.

Image from Stable Diffusion on OpenArt



Dog



Not Dog

Fourier analytic Barron Space theory

Discussion of Applications

- While your distribution doesn't need to be i.i.d. of data (in general it isn't), if it is very not i.i.d. it might be good to do some transformation first (PCA, VAE, etc).
- For shallow neural network $\sum_{j}^{M} a_j \sigma(\mathbf{b}_j \cdot \mathbf{x} + c_j)$, we can transform into Barron-E canonical form $\sum_{j}^{M} a_j \|\mathbf{b}_j\|_1 \sigma(\hat{\mathbf{b}}_j \cdot \mathbf{x} + c_j/\|\mathbf{b}_j\|_1)$ and then identify $a_j\|\mathbf{b}_j\|_1$ with $\rho(\omega_j, \beta_j, z_j)$, $\hat{\mathbf{b}}_j$ with $\hat{\omega}_j$ and $c_j/\|\mathbf{b}_j\|_1$ with $z_j\beta_j$.
- Obviously in practice there is another free parameter in a neural network (the outer weight) since z_j just identifies the sign. But in Barron-E canonical form we can consider the weights as being an interpolation of the coefficient.
  - Maybe better when quantized
- Other activation functions considered in theory, but just variations on ReLU. Same with multi-target.