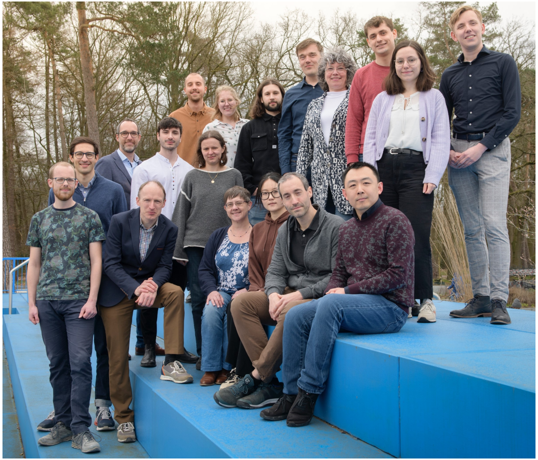
Duality for Neural Networks through Reproducing Kernel Banach Spaces

Len Spek

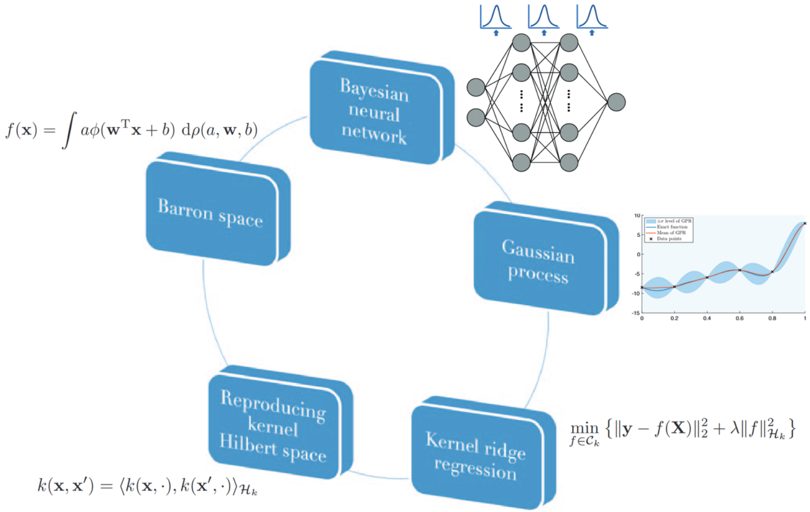
July 29, 2023

Mathematics of Imaging & AI @ University of Twente



Mathematics of Imaging & AI UNIVERSITY OF TWENTE.

Mathematics of Imaging & AI @ University of Twente



Basic Neural Network

Perceptron or shallow neural network with activation function $\sigma : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{1}{m} \sum_{j=1}^{m} a_j \sigma(\mathbf{v}_j^T x + b_j)$$

Basic Neural Network

Perceptron or shallow neural network with activation function $\sigma : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{1}{m} \sum_{j=1}^{m} a_j \sigma(\mathbf{v}_j^T x + b_j)$$

Universal approximation theorem: When $m \to \infty$, we can approximate any continuous function.

$$f(x) := A\pi = \int_\Omega a\sigma(\mathbf{v}^T x + b) d\pi(w)$$

π probability distribution of weights $w = (a, \mathbf{v}, b) \in \Omega$.

Basic Neural Network

Perceptron or shallow neural network with activation function $\sigma : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{1}{m} \sum_{j=1}^{m} a_j \sigma(\mathbf{v}_j^T x + b_j)$$

Universal approximation theorem: When $m \to \infty$, we can approximate any continuous function.

$$f(x) := A\pi = \int_{\Omega} a\sigma(\mathbf{v}^T x + b) d\pi(w)$$

π probability distribution of weights $w = (a, \mathbf{v}, b) \in \Omega$.

The functions f form a vector space. What norm?

Function Spaces for Neural Networks

Weinan E. and collaborators[1] introduced the Barron space.

$$f(x) = A\pi = \int_\Omega \sigma(v^T x + b) d\pi(w)$$

with the norm

$$\|f\| = \inf_{f=A\pi} \int_\Omega |a|(1 + \|v\|_1 + |b|) d\pi(w)$$

[1] E, Ma, and Wu, "A priori estimates of the population risk for two-layer neural networks".

[2] Parhi and Nowak, "Banach Space Representer Theorems for Neural Networks and Ridge Splines".

Function Spaces for Neural Networks

Weinan E. and collaborators[1] introduced the Barron space.

$$f(x) = A\pi = \int_\Omega \sigma(v^T x + b) d\pi(w)$$

with the norm

$$\|f\| = \inf_{f = A\pi} \int_\Omega |a|(1 + \|v\|_1 + |b|) d\pi(w)$$

For ReLu activation functions, Parhi and Nowak[2], define a normed space using the Radon transform based on ridge splines

$$\|f\| = c_d \|\partial_t^2 \Lambda^{d-1} \mathcal{R}\|$$

[1] E, Ma, and Wu, “A priori estimates of the population risk for two-layer neural networks”.

[2] Parhi and Nowak, “Banach Space Representer Theorems for Neural Networks and Ridge Splines”.

Challenges

- Connect different function spaces to reproducing kernel framework.

Challenges

- Connect different function spaces to reproducing kernel framework.
- Lack of an inner product, so no Hilbert space structure.

Challenges

- Connect different function spaces to reproducing kernel framework.
- Lack of an inner product, so no Hilbert space structure.
- Explore the dual structure of such function spaces? Does it help us understand the relation between data and weights

Reproducing Kernel Hilbert Spaces (RKHS)

Definition

Hilbert space $\mathcal{H}$ of functions $f : X \to \mathbb{R}$ is an RKHS if

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}}$$

for all $f \in \mathcal{H}$.

Reproducing Kernel Hilbert Spaces (RKHS)

Definition

Hilbert space $\mathcal{H}$ of functions $f : X \to \mathbb{R}$ is an RKHS if

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}}$$

for all $f \in \mathcal{H}$.

By Riesz representation theorem, there exist a symmetric kernel $K : X \times X \to \mathbb{R}$

$$f(x) = \langle K(x, \cdot), f \rangle$$

Reproducing Kernel Hilbert Spaces (RKHS)

Theorem

A Hilbert space $\mathcal{H}$ of functions on X satisfies the RKHS property if and only if there exists a Hilbert space Ψ and a map $\psi : X \mapsto \Psi$ such that

$$\begin{aligned} \mathcal{H} &= \Psi/\mathcal{N}(A) \\ \|f\|_{\mathcal{H}} &= \inf_{f=A\nu} \|\nu\|_{\Psi} \end{aligned} \tag{1}$$

where A maps ***features*** *in Ψ to functions on X and is defined as*

$$(A\nu)(x) = \langle \psi(x), \nu \rangle \tag{2}$$

for all $x \in X$ and $\nu \in \Psi$.

Reproducing Kernel Banach Spaces (RKBS)

Definition

A **Banach space** $\mathcal{B}$ of functions $f : X \to \mathbb{R}$ is an RKBS when

$$|f(x)| \leq C_x \|f\|_{\mathcal{B}}$$

for all $f \in \mathcal{B}$

For example: the space of continuous functions over X with max norm

Reproducing Kernel Banach Spaces (RKBS)

Theorem

*A **Banach space** $\mathcal{B}$ of functions on X satisfies the RKBS property if and only if there exists a **Banach space** Ψ and a map $\psi : X \mapsto \Psi^*$ such that*

$$\begin{aligned} \mathcal{B} &= \Psi / \mathcal{N}(A) \\ \|f\|_{\mathcal{B}} &= \inf_{f = A\nu} \|\nu\|_{\Psi} \end{aligned} \tag{3}$$

where the linear transformation A maps elements of the Banach space Ψ to functions on X and is defined as

$$(A\nu)(x) := \langle \psi(x), \nu \rangle \tag{4}$$

for all $x \in X$ and $\nu \in \Psi$.[a]

[a] Bartolucci et al., "Understanding neural networks with reproducing kernel Banach spaces".

A class of integral RKBS

Let $\mu \in \mathcal{M}(\Omega)$ a Radon measure and $\varphi \in C_0(X \times \Omega)$

$$f(x) := A\mu = \int_\Omega \varphi(x, w) d\mu(w)$$

Then we define the variational space $\mathcal{F}(X, \Omega)$[3] as

$$\mathcal{F}(X, \Omega) := \{f : X \to \mathbb{R} | \exists \mu \in \mathcal{M}(\Omega) \text{ s.t. } f = A\mu\}$$
$$\|f\| := \inf_{f = A\mu} \|\mu\|_{\mathcal{M}(\Omega)} = \inf_{f = A\mu} |\mu|(\Omega)$$

[3] Bach, "Breaking the Curse of Dimensionality with Convex Neural Networks".

[4] Bartolucci et al., "Understanding neural networks with reproducing kernel Banach spaces".

A class of integral RKBS

Let $\mu \in \mathcal{M}(\Omega)$ a Radon measure and $\varphi \in C_0(X \times \Omega)$

$$f(x) := A\mu = \int_\Omega \varphi(x,w)d\mu(w)$$

Then we define the variational space $\mathcal{F}(X,\Omega)$[3] as

$$\mathcal{F}(X,\Omega) := \{f : X \to \mathbb{R} | \exists \mu \in \mathcal{M}(\Omega) \text{ s.t. } f = A\mu\}$$
$$\|f\| := \inf_{f=A\mu} \|\mu\|_{\mathcal{M}(\Omega)} = \inf_{f=A\mu} |\mu|(\Omega)$$

Bartolucci and collaborators[4] showed that this RKBS admits a Representer Theorem and that the Radon regularisation is an instance of such an RKBS.

[3] Bach, "Breaking the Curse of Dimensionality with Convex Neural Networks".

[4] Bartolucci et al., "Understanding neural networks with reproducing kernel Banach spaces".

Barron Spaces and RKBS

We showed that Barron spaces also have an integral RKBS structure, where the Barron norm is equal to the variational norm.

If σ is 1-homogeneous, take $\Omega = \mathbb{S}^{d+1}$ and $w = (v, b)$

$$\varphi(x, w) = \sigma(v^T x + b)$$

If σ grows sublinearly, take $\Omega = \mathbb{R}^{d+1}$ and $w = (v, b)$

$$\varphi(x, w) = \frac{\sigma(v^T + b)}{1 + \|v\|_1 + |b|}$$

Reproducing **Kernel** Banach Spaces (RKBS)

Where is the kernel in RKBS?

Reproducing **Kernel** Banach Spaces (RKBS)

Where is the kernel in RKBS?

Challenge: No inner product $\implies$ No Riesz representation theorem

Reproducing **Kernel** Banach Spaces (RKBS)

Where is the kernel in RKBS?

Challenge: No inner product $\implies$ No Riesz representation theorem

Adjoint RKBS can be used to define a reproducing kernel. However, we lose symmetry of the kernel!

Adjoint RKBS

Definition

If the dual space $\mathcal{B}$ is as a space of functions on a set Ω and if there exists a function $K : X \times \Omega \to \mathbb{R}$, such that $K(x, \cdot) \in \mathcal{B}^*$ for all $x \in X$ and

$$f(x) = \langle K(x, \cdot), f \rangle$$

for all $x \in X$ and $f \in \mathcal{B}$, then we call K a **reproducing kernel** for $\mathcal{B}$.
If $\mathcal{B}^*$ is also an RKBS on Ω and it holds that $K(\cdot, w) \in \mathcal{B}$ for all $w \in \Omega$ and

$$g(w) = \langle g, K(\cdot, w) \rangle$$

for all $w \in \Omega$ and $g \in \mathcal{B}^*$, then we call $\mathcal{B}^*$ an **adjoint RKBS** of $\mathcal{B}$.
Then $K^*(w, x) := K(x, w)$ is a reproducing kernel of $\mathcal{B}^*$.[a]

[a] Lin, H. Z. Zhang, and J. Zhang, "On Reproducing Kernel Banach Spaces".

Adjoint Neural Networks Spaces

We define a new space $\mathcal{G}(X, \Omega)$ of 'Adjoint Neural Networks'.

Let $\rho \in \mathcal{M}(X)$ a Radon measure and $\varphi \in C_0(X \times \Omega)$

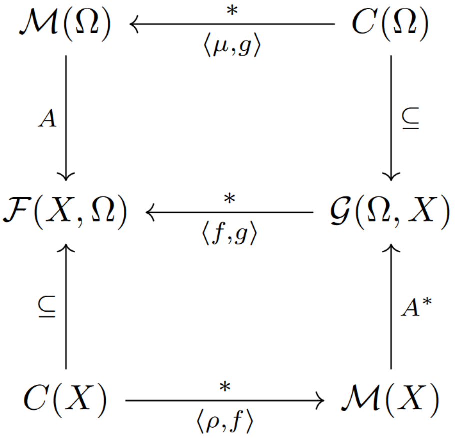
$$g(w) := A^* \rho = \int_X \varphi(x, w) d\rho(x)$$

Define the norm of g:

$$\mathcal{G}(X, \Omega) := \{g \in C_0(\Omega) | \exists \rho \in \mathcal{M}(X) \text{ s.t. } g = A^* \rho\}$$
$$\|g\|_{\mathcal{G}(X,\Omega)} := \sup_{w \in \Omega} |g(w)|$$

RKBS as point evaluation is bounded.

Duality diagram



Main Theorem: Data-Weight Duality

Theorem

*$\mathcal{F}(X, \Omega)$ is the **dual space** of $\mathcal{G}(\Omega, X)$ with the pairing*

$$\langle f, g\rangle := \langle \rho, f\rangle = \langle \mu, g\rangle = \langle \rho \times \mu, \varphi\rangle = \int_{X\times\Omega} \varphi(x, w) d(\rho \times \mu)(x, w)$$

where $f = A\mu$, $g = A^*\rho$.
*Furthermore, $\mathcal{F}(X, \Omega)$ and $\mathcal{G}(\Omega, X)$ form an adjoint pair of RKBS with **reproducing kernel** φ.*[a]

[a] Spek et al., *Duality for Neural Networks through Reproducing Kernel Banach Spaces*.

Proof Sketch - Pairing

First show that the duality pairing is well-defined using Fubini

$$\int_{X\times\Omega}\varphi(x,w)d(\rho\times\mu)(x,w)=\int_X\int_\Omega\varphi(x,w)d\mu(w)d\rho(x)=\int_X f(x)d\rho(x)=\langle\rho,f\rangle$$

$$\int_{X\times\Omega}\varphi(x,w)d(\rho\times\mu)(x,w)=\int_\Omega\int_X\varphi(x,w)d\rho(x)d\mu(w)=\int_\Omega g(w)d\mu(w)=\langle\mu,g\rangle$$

Hence **independent** of choice of μ and ρ

Proof Sketch - Pairing

First show that the duality pairing is well-defined using Fubini

$$\int_{X\times\Omega} \varphi(x,w)d(\rho\times\mu)(x,w) = \int_X \int_\Omega \varphi(x,w)d\mu(w)d\rho(x) = \int_X f(x)d\rho(x) = \langle \rho, f\rangle$$

$$\int_{X\times\Omega} \varphi(x,w)d(\rho\times\mu)(x,w) = \int_\Omega \int_X \varphi(x,w)d\rho(x)d\mu(w) = \int_\Omega g(w)d\mu(w) = \langle \mu, g\rangle$$

Hence **independent** of choice of μ and ρ

$$|\langle f, g\rangle| = |\langle \mu, g\rangle| \leq \|\mu\|_{\mathcal{M}(\Omega)}\|g\|_{C_0(\Omega)}$$

Taking the inifimum over μ s.t. $f = A\mu$

$$|\langle f, g\rangle| \leq \|f\|_{\mathcal{F}(X,\Omega)}\|g\|_{\mathcal{G}(\Omega,X)}$$

Proof Sketch - Duality

As $\mathcal{F}(X, \Omega)$ is a quotient space

$$\mathcal{F}(X, \Omega) := \mathcal{M}(\Omega)/\mathcal{N}(A)$$

Its dual is given by the annihilator of $\mathcal{N}(A)$, i.e. all $g \in C_0(\Omega)$ s.t.

$$\langle \mu, g \rangle = 0$$

for all μ s.t. $A\mu = 0$.

Proof Sketch - Duality

As $\mathcal{F}(X,\Omega)$ is a quotient space

$$\mathcal{F}(X,\Omega) := \mathcal{M}(\Omega)/\mathcal{N}(A)$$

Its dual is given by the annihilator of $\mathcal{N}(A)$, i.e. all $g \in C_0(\Omega)$ s.t.

$$\langle \mu, g \rangle = 0$$

for all μ s.t. $A\mu = 0$. This turns out to be exactly the space $\mathcal{G}(\Omega, X)$ as

$$\langle \mu, g \rangle = \langle \rho, A\mu \rangle = 0$$

for some $\rho \in \mathcal{M}(X)$ s.t. $g = A^*\rho$.

Proof Sketch - Reproducing Kernel

To show that φ is indeed the Reproducing Kernel

$$f(x) = \langle f, \varphi(x,\cdot)\rangle \quad \text{and} \quad g(w) = \langle \varphi(x,\cdot), g\rangle$$

Proof Sketch - Reproducing Kernel

To show that φ is indeed the Reproducing Kernel

$$f(x) = \langle f, \varphi(x, \cdot) \rangle \quad \text{and} \quad g(w) = \langle \varphi(x, \cdot), g \rangle$$

We use that

$$\varphi(x, \cdot) = \int_X \varphi(x', \cdot) d\delta_x(x') = A^* \delta_x \in \mathcal{G}(\Omega, X)$$

And by the duality pairing

$$\langle f, \varphi(x, \cdot) \rangle = \langle f, \delta_x \rangle = f(x)$$

Looking Forward

Using this dual framework, we have derived the dual problem and shown strong duality.

Looking Forward

Using this dual framework, we have derived the dual problem and shown strong duality. Estimation error - Approximation error Duality

Looking Forward

Using this dual framework, we have derived the dual problem and shown strong duality. Estimation error - Approximation error Duality

Leveraging duality in optimisation: Use in experimental design or architecture search.

Example: Bregman iteration for neural networks and optimality conditions. The duality can be used to determine a source condition which lives in the dual space.

Looking Forward

Using this dual framework, we have derived the dual problem and shown strong duality. Estimation error - Approximation error Duality

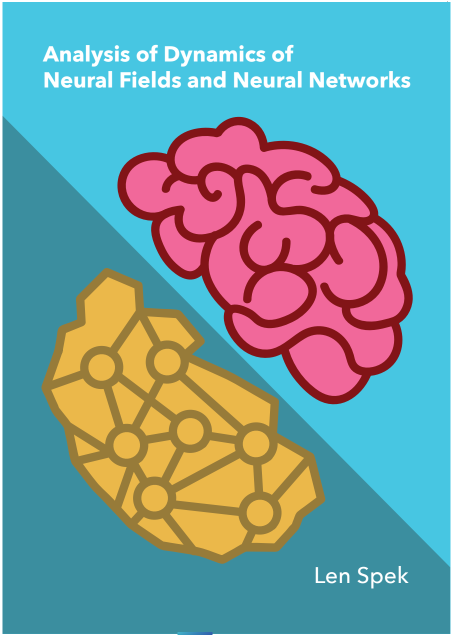
Leveraging duality in optimisation: Use in experimental design or architecture search.

Example: Bregman iteration for neural networks and optimality conditions. The duality can be used to determine a source condition which lives in the dual space.

Further goals: Expanding RKBS to deep networks and exploring the role of depth.

Thank you for your attention

Len Spek, Tjeerd Jan Heeringa, Felix Schwenninger, Christoph Brune.
"Duality for neural networks through reproducing kernel Banach spaces."
arXiv preprint arXiv:2211.05020 (2023).



Dual formulation of ERM

Primal problem: Given a target $y : X \to \mathbb{R}$ and a data distribution $\nu \in \mathcal{M}(X)$

$$\inf_{\mu \in \mathcal{M}(\Omega)} \frac{1}{2} \|A\mu - y\|^2_{L^2(\nu)} + |\mu|(\Omega)$$

Dual problem:

$$\sup_{\rho \in \mathcal{M}(X)} -J^*(-\rho) - R^*(A^*\rho)$$

$$J^*(\rho) = \begin{cases} \int_X \frac{1}{2} \frac{d\rho}{d\nu}(x) + y(x) d\rho(x) & \rho \ll \nu \\ \infty & \text{otherwise} \end{cases}$$
$$R^*(g) = \begin{cases} 0 & \|g\|_{C_0(\Omega)} \leq 1 \\ \infty & \text{otherwise} \end{cases} \tag{5}$$