A Banach Space Representer Theorem for Single-Hidden Layer Neural Networks

Rahul Parhi
Department of Electrical and Computer Engineering
University of Wisconsin–Madison

(joint work with Robert Nowak)

SLowDNN
November 24th, 2020

What is a representer theorem?

Definition

A **representer theorem** designates a **finite-dimensional parametric formulation** of solutions to a learning problem posed in a possibly **infinite-dimensional** space, ideally being a linear combination from some dictionary of atoms.

Classical representer theorems

- First studied in the context of smoothing splines in $H^k(\mathbb{R})$.

  $\Longrightarrow$ Kimeldorf & Wahba (1970, 1971)
- Later studied in the general setting of reproducing kernel Hilbert spaces.

  $\Longrightarrow$ Wahba (1990)

Classical representer theorems

Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel $k(\cdot,\cdot)$ and consider the scattered data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$. Then,

$$\min_{f \in \mathcal{H}} \sum_{n=1}^{N} \ell(f(\boldsymbol{x}_n), y_n) + \lambda \|f\|_{\mathcal{H}}, \quad \lambda > 0, \tag{1}$$

admits a solution f^* of the form $f^*(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n \, k(\boldsymbol{x}, \boldsymbol{x}_n)$.

$\Longrightarrow$ can simply optimize over $\{\alpha_n\}_{n=1}^N$ to solve (1).

Modern representer theorems

Moving beyond Hilbert spaces:

- Recently, the term "representer theorem" started being used for more general problems about convex regularization.
  - $\Longrightarrow$ Unser et al. (2017) – Banach spaces
  - $\Longrightarrow$ Boyer et al. (2019) – locally convex spaces
  - $\Longrightarrow$ Bredies & Carioni (2020) – locally convex spaces
- Reproducing Kernel Banach Spaces
  - $\Longrightarrow$ Zhang et al. (2009)
  - $\Longrightarrow$ Xu & Ye (2019)
- Many classical results in Banach spaces
  - $\Longrightarrow$ Zuhovickiĭ (1948) – Radon measure recovery
  - $\Longrightarrow$ Fisher & Jerome (1975) – Radon measure recovery, L^1 splines
  - $\Longrightarrow$ Mammen & van de Geer (1997) – Locally adaptive regression splines

Neural network representer theorem

Question

Is there a representer theorem for (single-hidden layer) neural networks?

Answer

Yes! But in a non-Hilbertian Banach space.

Neural network representer theorem

Theorem (P. & Nowak, 2020)

There is a family of Banach spaces $\mathcal{F}_m$ and family of seminorms $\|\cdot\|_{(m)}$ such that for any scattered data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$, there **exists** a solution f^* to

$$\min_{f \in \mathcal{F}_m} \sum_{n=1}^{N} \ell(f(\boldsymbol{x}_n), y_n) + \lambda \|f\|_{(m)}, \quad \lambda > 0, \tag{2}$$

of the form

$$f^*(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \rho_m(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k) + c(\boldsymbol{x}), \quad K < N.$$

$\Longrightarrow$ can simply optimize over $\{v_k, \boldsymbol{w}_k, b_k\}_{k=1}^K$ and c to solve (2).

Neural network representer theorem

- $\|f\|_{(m)} := \left\|\partial_t^m \Lambda^{d-1} \mathscr{R} f\right\|_{\mathcal{M}}$

  $\Longrightarrow$ $\mathscr{R}$ – Radon transform

  $\Longrightarrow$ Λ^{d-1} – Ramp filter

  $\Longrightarrow$ ∂_t^m – m partial derivatives in offset variable of Radon domain

  $\Longrightarrow$ $\|\cdot\|_{\mathcal{M}}$ – TV norm (in the sense of measures). $L^1 \subset \mathcal{M} \subset \mathscr{S}'$, but $\mathcal{M}$ includes distributions such as the Dirac impulse.

- $\mathcal{F}_m := \left\{f : \mathbb{R}^d \to \mathbb{R} : \left\|\partial_t^m \Lambda^{d-1} \mathscr{R} f\right\|_{\mathcal{M}} < \infty\right\}$

- $\rho_m = \max\{0, \cdot\}^{m-1}/(m-1)!$ – truncated power functions

  $\Longrightarrow$ $m = 2$ corresponds to ReLU networks.

- c is a "generalized bias" term, i.e., a polynomial of degree $< m$.

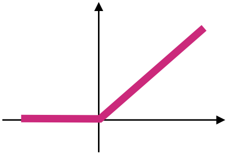
Why the Radon transform?

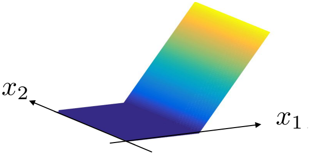
- The Radon transform computes integrals over **hyperplanes**.

$\implies$ $\mathscr{R}\{f\}(\boldsymbol{\gamma}, t) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \delta(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t) \, \mathrm{d}\boldsymbol{x}$

$\implies$ Radon domain parameterized by a **direction** γ and an **offset** t.

- Single-hidden layer neural networks are superpositions of **ridge functions**.





- A neuron is a mapping of the form $\boldsymbol{x} \mapsto \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)$.

$\implies$ Parameterized by a **direction** $\boldsymbol{w}$ and an **offset** b.

$\implies$ The Radon transform provides a convenient way to "extract" the direction and offset from a neuron.

$\implies$ $\partial_t^m \Lambda^{d-1} \mathscr{R}\left\{\rho_m(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\right\} = \delta_{\mathsf{Radon}}(\cdot - (\boldsymbol{w}, b))$.

Radon transform and ridge functions

pre-1950s: Superpositions of plane waves are solutions to many PDEs, e.g., the wave equation.

$\implies$ Plane waves are just ridge functions.

$\implies$ Radon domain analysis is useful.

1970s: Seminal paper on **computerized tomography** from Logan & Shepp (1975).

$\implies$ Coined the term "ridge function".

1990s: Multiscale system referred to as **ridgelets** proposed by Murata (1996); Rubin (1998); Candès (1998, 1999).

$\implies$ Ridglet transform is just a one-dimensional wavelet transform in the Radon domain.

2020: Ongie et al. (2020) show that $\left\|\partial_t^2 \Lambda^{d-1} \mathscr{R} f\right\|_{\mathcal{M}}$ captures the Euclidean norm of the weights in an infinite-width ReLU network.

$\implies$ Provides insight into what functions can be **represented** by infinite-width ReLU networks.

Remarks

- Much of past work has focused on characterizing what functions can be **approximated** or **represented** by neural networks.

$\Longrightarrow$ Not practically interesting.

- The utility of our representer theorem says what happens when one **trains** a neural network on **data**.

Finite-dimensional neural network training

- The utility of RKHS representer theorems is that the infinite-dimensional optimizations can be recast as finite-dimensional optimizations.
- Also applies to our neural network representer theorem.

$$\implies \quad \text{Let } f_{\boldsymbol{\theta}}(\boldsymbol{x}) := \sum_{k=1}^{K} v_k \, \rho_m(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + c(\boldsymbol{x}).$$

$$\implies \quad \left\| \partial_t^m \Lambda^{d-1} \mathscr{R} \, f_{\boldsymbol{\theta}} \right\|_{\mathcal{M}} = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1}$$

Finite-dimensional neural network training

- Can consider the finite-dimensional optimization

$$\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n), y_n) + \lambda \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1}$$

$\Longrightarrow$ A kind of path-norm regularization (Neyshabur et al., 2015).

- Which is equivalent to

$$\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n), y_n) + \lambda \sum_{k=1}^{K} |v_k|^2 + \|\boldsymbol{w}_k\|_2^{2m-2}$$

$\Longrightarrow$ A kind of weight decay regularization (Krogh & Hertz, 1992).

Takeaway messages

- Representer theorems are much more general than the well-known RKHS setting.
- **Nonparametric** learning problems with $\left\|\partial_t^m \Lambda^{d-1} \mathscr{R}\{\cdot\}\right\|_{\mathcal{M}}$-norm regularization have **sparse, atomic solutions** which are single-hidden layer neural networks.
- $\left\|\partial_t^m \Lambda^{d-1} \mathscr{R}\{\cdot\}\right\|_{\mathcal{M}}$ is equivalent to neural network path-norms.
- $\left\|\partial_t^m \Lambda^{d-1} \mathscr{R}\{\cdot\}\right\|_{\mathcal{M}}$-norm regularization is equivalent to forms of weight decay.
- Regularizers are "matched" to the activation function.

Thank You

Questions?

https://arxiv.org/abs/2006.05626

References

Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2): 1260–1281, 2019.

Kristian Bredies and Marcello Carioni. Sparsity of solutions for variational inverse problems with finite-dimensional data. *Calculus of Variations and Partial Differential Equations*, 59(1):14, 2020.

Emmanuel J. Candès. *Ridgelets: theory and applications*. PhD thesis, Stanford University Stanford, 1998.

Emmanuel J. Candès. Harmonic analysis of neural networks. *Applied and Computational Harmonic Analysis*, 6(2):197–218, 1999.

S. D. Fisher and Joseph W. Jerome. Spline solutions to L^1 extremal problems in one and several variables. *Journal of Approximation Theory*, 13(1):73–83, 1975.

References

George S. Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

George S. Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pp. 950–957, 1992.

Benjamin F. Logan and Larry A. Shepp. Optimal reconstruction of a function from its projections. *Duke mathematical journal*, 42(4): 645–659, 1975.

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

References

Noboru Murata. An integral representation of functions using three-layered networks and their approximation bounds. *Neural Networks*, 9(6):947–956, 1996.

Behnam Neyshabur, Russ R. Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2015.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Boris Rubin. The Calderón reproducing formula, windowed X-ray transforms, and Radon transforms in L^p-spaces. *Journal of Fourier Analysis and Applications*, 4(2):175–197, 1998.

References

Michael Unser, Julien Fageot, and John Paul Ward. Splines are universal solutions of linear inverse problems with generalized TV regularization. *SIAM Review*, 59(4):769–793, 2017.

Grace Wahba. *Spline models for observational data*, volume 59. SIAM, 1990.

Yuesheng Xu and Qi Ye. *Generalized Mercer kernels and reproducing kernel Banach spaces*, volume 258. American Mathematical Society, 2019.

Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, 10(Dec):2741–2775, 2009.

S. Zuhovickiĭ. Remarks on problems in approximation theory. *Mat. Zbirnik KDU*, pp. 169–183, 1948.