Deep Spline Network With Control of Lipschitz Regularity

Shayan Aziznejad   Michael Unser

Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne, Switzerland.

Supervised Learning

Training Dataset $(\boldsymbol{x}_m, y_m) \in \mathbb{R}^{d+1}$

Objective Function $f : \mathbb{R}^d \to \mathbb{R}$, $y_m \approx f(\boldsymbol{x}_m)$.

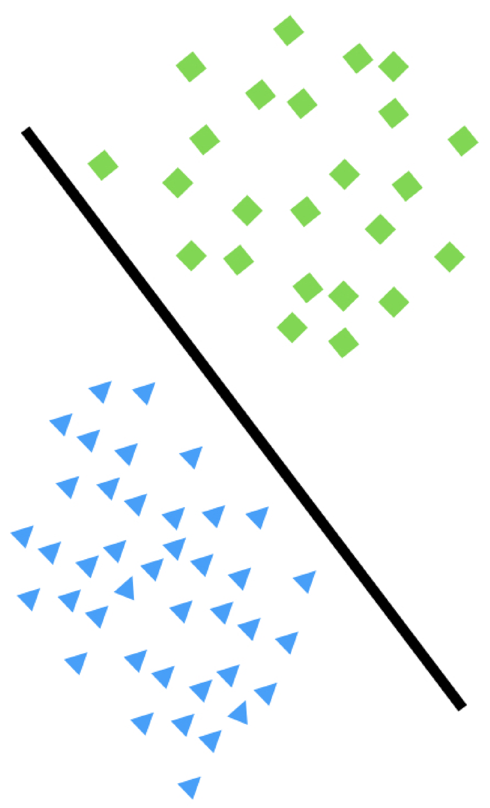
Formulation Through an optimization problem

$$\min_{f \in \mathcal{F}} \sum_{m=1}^{M} \mathrm{E}\left(y_m, f(\boldsymbol{x}_m)\right) + \lambda \mathcal{R}(f), \tag{1}$$

- $\mathrm{E}(\tilde{y}, y)$: The error function,
- $\mathcal{R}(f)$: The regularization,
- $\mathcal{F}$: The search space

Example Ridge Regression

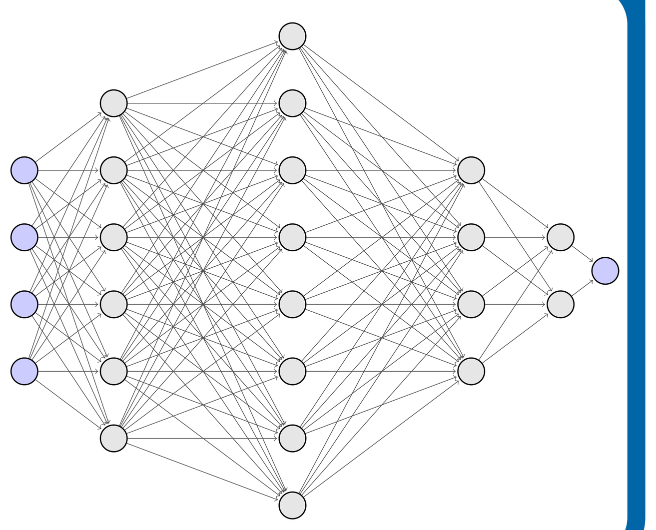
$$\min_{\boldsymbol{a} \in \mathbb{R}^N} \sum_{m=1}^{M} (y_m, \boldsymbol{a}^T \boldsymbol{x}_m)^2 + \lambda \|\boldsymbol{a}\|^2. \tag{2}$$



Deep Learning Model

Composition of parametric vector-valued functions

- $\mathbf{f}_{\text{deep}} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L} : \boldsymbol{x} \mapsto \mathbf{f}_L \circ \cdots \circ \mathbf{f}_1(\boldsymbol{x})$
- $\mathbf{f}_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$: the ℓth layer
- The nth neuron of $\mathbf{f}_\ell$: $\boldsymbol{x} \mapsto \sigma_{n,\ell}(\mathbf{w}_{n,\ell}^T \boldsymbol{x})$,
  - $\mathbf{w}_{n,\ell} \in \mathbb{R}^{N_{\ell-1}}$ are linear weights and,
  - $\sigma_{n,\ell} : \mathbb{R} \to \mathbb{R}$ are point-wise nonlinearities.



Activation Functions

Standard Paradigm

Fix the shape of neurons

- $\sigma_{n,\ell}(x) = \sigma(x - b_{n,\ell})$
- Learn the bias terms $b_{n,\ell}$
- Example: Rectified Linear Unit (ReLU) [4]

Learning Parametric Activations

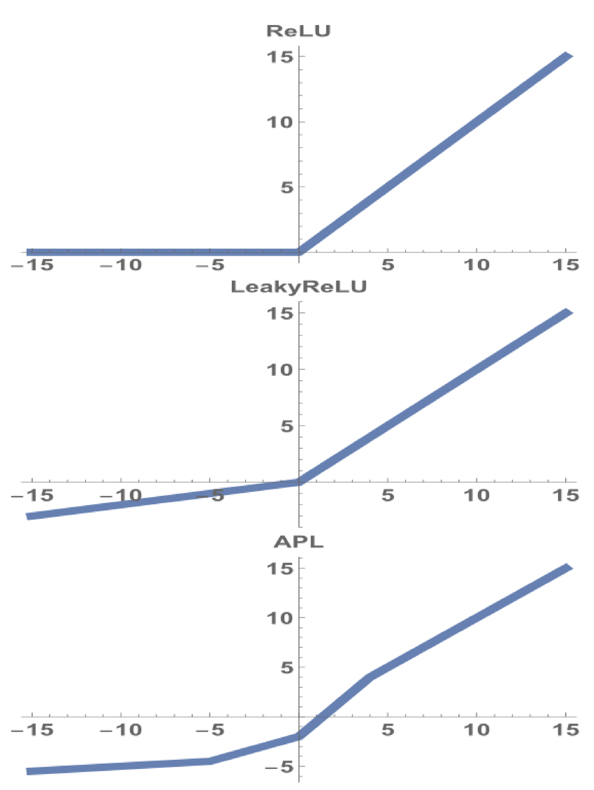
- Adaptive Leaky ReLU [3]
- Adaptive piece-wise linear [1]

Our Proposal Variational Formulation (1) [6]

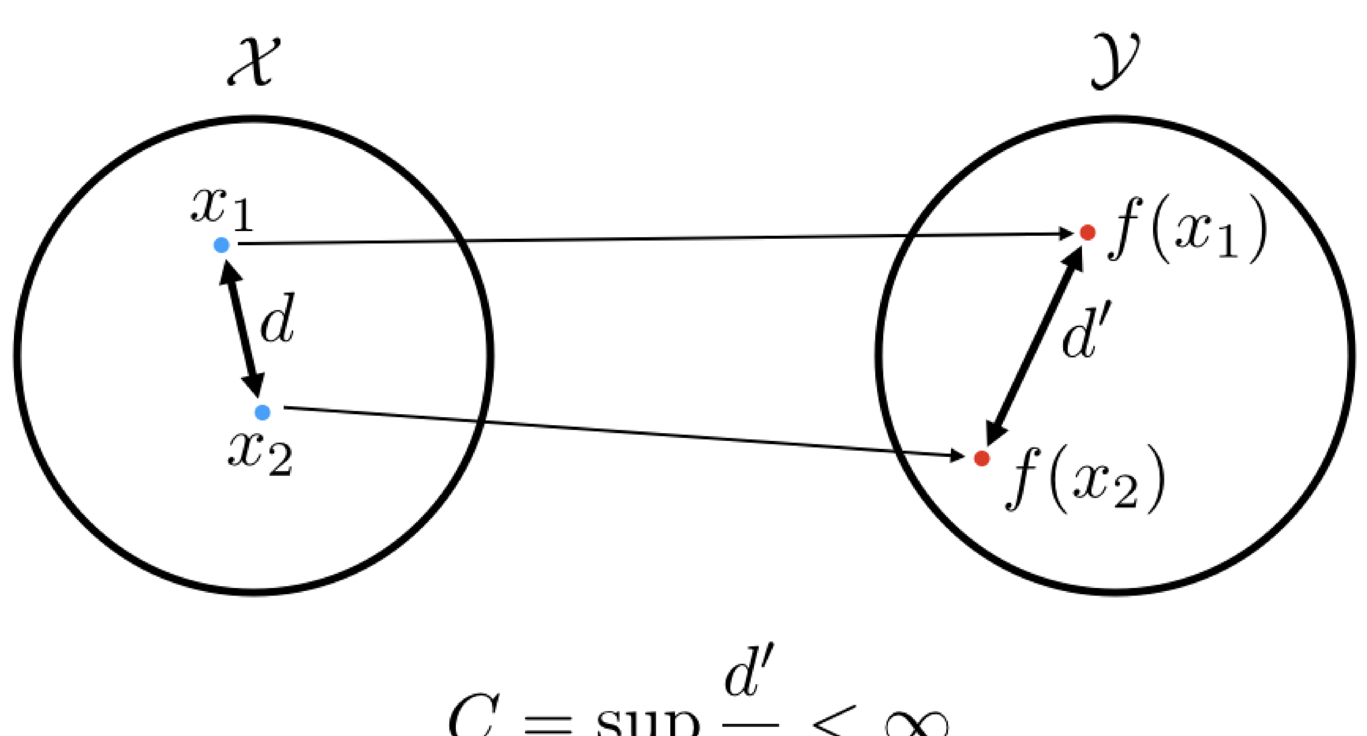
- The search space: $\mathcal{F} = \mathrm{BV}^{(2)}(\mathbb{R})$
- The regularization: $\mathcal{R}(\sigma) = \|\sigma\|_{\mathrm{BV}^{(2)}}$

Regarding the search space $\mathrm{BV}^{(2)}(\mathbb{R})$:

- $\mathrm{BV}^{(2)}(\mathbb{R}) = \{\sigma : \mathbb{R} \to \mathbb{R} : \|\sigma\|_{\mathrm{BV}^{(2)}} < \infty\}$,
- $\|\sigma\|_{\mathrm{BV}^{(2)}} = \|\mathrm{D}^2\sigma\|_{\mathcal{M}} + |\sigma(0)| + |\sigma(1) - \sigma(0)|$,



Lipschitz Regularity



$$C = \sup \frac{d'}{d} < \infty$$

Definition 1 *A function $f : \mathcal{X} \to \mathcal{Y}$ ($\mathcal{X}$ and $\mathcal{Y}$ are normed spaces with their corresponding norms denoted by $\|\cdot\|_{\mathcal{X}}$, $\|\cdot\|_{\mathcal{Y}}$, respectively) is Lipschitz if, for all $x_1, x_2 \in \mathcal{X}$, there exists a constant C such that*

$$\boxed{\|f(x_1) - f(x_2)\|_{\mathcal{Y}} \leq C\|x_1 - x_2\|_{\mathcal{X}}.} \tag{3}$$

Proposition 1 *Any function $\sigma \in \mathrm{BV}^{(2)}(\mathbb{R})$ is Lipschitz-continuous with constant $C = \|\sigma\|_{\mathrm{BV}^{(2)}}$.*

Why Lipschitz?

- Any Lipschitz function is continuous and almost everywhere differentiable.
  $\Rightarrow$ essential for back-propagation
- Generalization property of deep neural networks [5]
- Convergence analysis in deep learning schemes [2]

Global Lipschitz Bound

Theorem 1 (Lipschitz regularity of deep neural networks) *Any feed-forward fully-connected deep neural network with the nonlinearity selected from the space $\mathrm{BV}^{(2)}(\mathbb{R})$ and normalized linear weights (with respect to the ℓ_∞-norm) specifies an input-output relation that is Lipschitz with respect to the ℓ_1-norm with constant*

$$\boxed{C = \prod_{\ell=1}^{L} \left(\sum_{n=1}^{N_\ell} \|\sigma_{n,\ell}\|_{\mathrm{BV}^{(2)}} \right).} \tag{4}$$

- Optimizing $\sum_{n=1}^{N_\ell} \|\sigma_{n,\ell}\|_{\mathrm{BV}^{(2)}}$ contributes to a decrease of the overall Lipschitz constant of the network.
  $\Rightarrow$ Motivation for including these terms in the regularization functional

Problem Formulation

We formulate our training problem as

$$\boxed{\min_{\substack{\|\mathbf{w}_{n,\ell}\|_\infty = 1 \\ \sigma_{n,\ell} \in \mathrm{BV}^{(2)}(\mathbb{R})}} \sum_{m=1}^{M} \mathrm{E}\left(\mathbf{y}_m, \mathbf{f}(\boldsymbol{x}_m)\right) + \mu \sum_{\ell=1}^{L} \sum_{n=1}^{N_\ell} R_\ell(\mathbf{w}_{n,\ell}) + \lambda \sum_{\ell=1,}^{L} \left(\sum_{n=1}^{N_\ell} \|\sigma_{n,\ell}\|_{\mathrm{BV}^{(2)}} \right),} \tag{5}$$

where
- $\mathrm{E}(\tilde{\mathbf{y}}, \mathbf{y})$: The error function with $\mathrm{E}(\mathbf{y}, \mathbf{y}) = 0$,
- $R_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}_{\geq 0}$: Regularization term for linear weights,
- $\lambda, \mu \in \mathbb{R}_{>0}$: Adjustable parameters.

Representer Theorem

Theorem 2 ($\mathrm{BV}^{(2)}$ optimality of deep splines) *If the solution of (5) exists, then it is achieved by a deep spline network with individual activations of the form*

$$\boxed{\sigma_{n,\ell}(x) = b_{1,n,\ell} + b_{2,n,\ell}x + \sum_{k=1}^{K_{n,\ell}} a_{k,n,\ell}\mathrm{ReLU}(x - \tau_{k,n,\ell}),} \tag{6}$$

with adaptive parameters $K_{n,\ell} \leq M$, $\tau_{1,n,\ell}, \ldots, \tau_{K_{n,\ell},n,\ell} \in \mathbb{R}$, and $b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \ldots, a_{K_{n,\ell},n,\ell} \in \mathbb{R}$.

- The classical ReLU networks are special cases of our solution form.
- The $\mathrm{BV}^{(2)}$-norm of the activations
  $\|\sigma_{n,\ell}\|_{\mathrm{BV}^{(2)}} = \sum_{k=1}^{K_{n,\ell}} |a_{k,n,\ell}| + |b_1| + |b_2|$
  $\Longrightarrow$ ℓ_1 minimization techniques.

References

[1] Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830*, 2014.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[4] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[5] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[6] Michael Unser. A representer theorem for deep neural networks. *arXiv preprint arXiv:1802.09210*, 2018.

Acknowledgment This work was funded by the Swiss National Science Foundation under Grant 200020_162343 / 1