DEEP RITZ REVISITED

Johannes Müller
Max Planck Institut for Mathematics in the Sciences
jmueller@mis.mpg.de

Marius Zeinhofer
University of Freiburg
marius.zeinhofer@mathematik.uni-freiburg.de

ABSTRACT

Recently, progress has been made in the application of neural networks to the numerical analysis of stationary and instationary partial differential equations. For example, one can use the variational formulation of the Dirichlet problem in order to obtain an objective function – a penalised Dirichlet energy – for the optimization of the parameters of neural networks with a fixed architecture. Although this approach yields promising empirical results especially in high dimensions it is lacking any convergence guarantees. We use the notion of Γ-convergence to show that ReLU networks of growing architecture that are trained with respect to suitably penalised Dirichlet energies converge to the solution of the Dirichlet problem. We discuss how our findings generalise to arbitrary variational problems under certain universality assumptions on the neural networks that are used. We see that this covers nonlinear stationary PDEs like the p-Laplace.

1 INTRODUCTION

Artificial neural networks play a key role in current machine learning research and both their performance in practice as well as numerous of their theoretical properties are studied extensively. After their initial success in supervised learning problems (see Krizhevsky et al., 2012) neural networks were successfully used in a variety of fields like generative and reinforcement learning (see Goodfellow et al., 2014; Mnih et al., 2013). More recently they have also been applied to inverse problems and the numerical analysis of PDEs (see McCann et al., 2017; E et al., 2017, respectively). Some of the deep learning based approaches to their numerical solution are similar to traditional learning problems. In fact, the Feynman-Kac formula allows an approximate evaluation of the solution of a Kolmogorov PDE via Monte-Carlo sampling and neural networks can be used for regression of those evaluations (see Berner et al., 2018). However, one can also use the variational formulation of elliptic PDEs and use the resulting energy as an objective function for the optimisation of the parameters of the neural network.

More precisely, let us consider the Dirichlet problem on $\Omega \subseteq \mathbb{R}^d$ which is the prototype of an elliptic PDE and given by

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1}$$

It is well known that for a function $u \in H_0^1(\Omega)$[1] and a square integrable right hand side $f \in L^2(\Omega)$ it is equivalent to be a weak solution[2] of the Dirichlet problem or to be a solution of the variational problem

$$u \in \operatorname*{arg\,min}_{v \in H_0^1(\Omega)} \int_\Omega \left(\frac{1}{2} |\nabla v|^2 - fv \right) \mathrm{d}x.$$

The objective function on the right hand side is known as the *Dirichlet energy*. In the remainder we interpret neural networks as real functions on $\Omega \subseteq \mathbb{R}^d$ which are elements of the space $H^1(\Omega)$ for sufficiently regular activation functions. Now, one could minimise the Dirichlet energy over all neural networks with zero boundary values, i.e., all networks in $H_0^1(\Omega)$. Since the Dirichlet energy is

[1] The space of square integrable functions with square integrable first weak derivatives is denoted by $H^1(\Omega)$, the subspace of functions with zero boundary values with $H_0^1(\Omega)$; see Brezis (2010) for further details.

[2] See Brezis (2010) for a short introduction to the concept of weak solutions.

coercive and continuous those trained networks would converge towards the solution of the Dirichlet problem if they are universal approximators. However, contrary to finite element functions it is not straight forward to enforce zero boundary value conditions for neural networks which makes this optimisation problem infeasible. The solution is to relax the problem and consider arbitrary networks but to penalise boundary values. This approach shows promising performance especially in high dimensions (see E and Yu, 2018). Due to the early work of Ritz (1909) on the finite dimensional approximation of variational problems this approach is known as the *deep Ritz method*. So far, however, this numerical framework for variational problems is lacking convergence guarantees.

CONTRIBUTIONS AND MAIN RESULT

We show that neural networks of growing architecture that are trained with respect to suitably penalised Dirichlet energies converge to the solution of the Dirichlet problem (1). This is our main contribution and proves the consistency of the deep Ritz method proposed by E and Yu (2018).

More precisely, let $(\Theta_n)_{n\in\mathbb{N}}$ be sets of parameters of neural networks that we assume to be universal approximators in $H_0^1(\Omega)$ for $n\to\infty$. Now we introduce the objective functions

$$L_n\colon \Theta_n \to \mathbb{R}, \quad \theta \mapsto \int_\Omega \left(\frac{1}{2}\left|\nabla u_\theta\right|^2 - f u_\theta\right)\mathrm{d}x + n\int_{\partial\Omega} u_\theta^2 \mathrm{d}s, \tag{2}$$

where u_θ is the network arising from the parameters θ and $f\in L^2(\Omega)$ is some right hand side.

Theorem 1. *Let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence of quasi-minimisers of the objective functions, meaning*

$$L_n(\theta_n) \le \inf_{\theta\in\Theta_n} L_n(\theta) + \delta_n,$$

where $\delta_n\to 0$. Then $(u_{\theta_n})_{n\in\mathbb{N}}$ converges to the solution u of the Dirichlet problem (1)*, both weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$.*

The setting including the requirements on the domain Ω and the sets Θ_n are presented in full detail later. Furthermore, this convergence result holds for a broad class of variational problems and the penalisation strengths n can be replaced by arbitrary λ_n that approach $+\infty$. In the following we restrict ourselves to the Dirichlet problem and postpone technical details to the appendix.

2 PRELIMINARIES FROM NEURAL NETWORK THEORY

Let for the remainder $d, m, L, N_0, \dots, N_L$ be natural numbers and let

$$\theta = ((A_1, b_1), \dots, (A_L, b_L))$$

be a tupel of matrix-vector pairs where $A_l\in\mathbb{R}^{N_l\times N_{l-1}}$ and $b_l\in\mathbb{R}^{N_l}$ and $N_0=d, N_L=m$. Every matrix vector pair (A_l, b_l) induces an affine linear transformation denoted by $T_l\colon \mathbb{R}^{N_{l-1}}\to\mathbb{R}^{N_l}$. The *neural network with parameters* θ and with respect to some *activation function* $\rho\colon\mathbb{R}\to\mathbb{R}$ is the function

$$u = u_\theta\colon \mathbb{R}^d\to\mathbb{R}^m, \quad x\mapsto T_L(\rho(T_{L-1}(\rho(\cdots\rho(T_1(x)))))).$$

If we have $f=u_\theta$ for some θ we say the function f is *expressed* by the neural network. We call d the *input* and m the *output dimension*, L the *depth* and $W(\theta) := \max_{l=0,\dots,L} N_l$ the *width* of the neural network. In the remainder we will restrict ourselves to the case $m=1$ since we only consider real valued functions.

Further, we restrict ourselves to a specific activation function which is not only widely used in practice (see Ramachandran et al., 2017) but also exhibits nice theoretical properties (see Arora et al., 2016; Petersen et al., 2018). The *rectified linear unit* or *ReLU activation function* is defined via $x\mapsto\max\{0,x\}$ and we call networks with this particular activation function *ReLU networks*. The class of ReLU networks coincides with the class of continuous and piecewise linear functions (see Arora et al., 2016). Since piecewise linear functions are dense in $H_0^1(\Omega)$ we obtain the following universal approximation result which we prove in detail in the appendix.

Theorem 2 (Universal approximation)**.** *Let $\Omega\subseteq\mathbb{R}^d$ be an open set and $u\in H_0^1(\Omega)$. Then for all $\varepsilon>0$ there exists a function $u_\varepsilon\in H_0^1(\Omega)$ that can be expressed by a ReLU network of depth $\lceil\log_2(d+1)\rceil+1$ such that*

$$\|u-u_\varepsilon\|_{H^1(\Omega)} < \varepsilon.$$

3 Γ-CONVERGENCE

We recall the definition of Γ-convergence, which we specialise to the case of reflexive spaces with their associated weak topologies. For further reading we point the reader towards Dal Maso (2012).

Definition 3 (Γ-convergence)**.** Let X be a reflexive Banach space, $F_n, F\colon X \to (-\infty, \infty]$. Then $(F_n)_{n\in\mathbb{N}}$ is called Γ-*convergent* to F if the following two properties are satisfied.

1. *Liminf inequality:* For every $x \in X$ and $(x_n)_{n\in\mathbb{N}}$ with $x_n \rightharpoonup x$[3] we have
$$F(x) \le \liminf_{n\to\infty} F_n(x_n).$$
2. *Recovery sequence:* For every $x \in X$ there is $(x_n)_{n\in\mathbb{N}}$ with $x_n \rightharpoonup x$ such that
$$F(x) = \lim_{n\to\infty} F_n(x_n).$$

The sequence $(F_n)_{n\in\mathbb{N}}$ is called *equi-coercive* if the set $\{x \in X \mid F_n(x) \le r \text{ for some } n\}$ is bounded in X (or equivalently relatively compact with respect to the weak topology) for all $r \in \mathbb{R}$. We say that a sequence $(x_n)_{n\in\mathbb{N}}$ are *quasi minimisers* of the functionals $(F_n)_{n\in\mathbb{N}}$ if we have $F_n(x_n) \le \inf_{x\in X} F_n(x) + \delta_n$ where $\delta_n \to 0$.

Our main result uses the following property of Γ-convergent sequences. We want to emphasise the fact that there are no requirements regarding the continuity of any of the functionals and that the functionals $(F_n)_{n\in\mathbb{N}}$ are not assumed to admit minimisers.

Theorem 4 (Convergence of quasi-minimisers)**.** *Let X be a reflexive Banach space and $(F_n)_{n\in\mathbb{N}}$ be an equi-coercive sequence of functionals that Γ-converges to F which has a unique minimiser x. Then every sequence $(x_n)_{n\in\mathbb{N}}$ of quasi-minimisers of the sequence $(F_n)_{n\in\mathbb{N}}$ converges weakly to x.*

4 OUTLINE OF THE PROOF

Let $d \in \mathbb{N}$ and let $\Omega \subseteq \mathbb{R}^d$ be an open, connected and bounded set with Lipschitz boundary $\partial\Omega$. For $n \in \mathbb{N}$ let Θ_n denote the set of parameters of networks with input dimension d, output dimension 1, depth $\lceil \log_2(d+1) \rceil + 1$ and width n. Further, we denote the set of ReLU networks arising from those parameters by $A_n := \{u_\theta \mid \theta \in \Theta_n\}$. Since ReLU networks are continuous and piecewise linear we have $A_n \subseteq H^1(\Omega)$. Now we fix a right hand side $f \in H^1(\Omega)'$, that is $f : H^1(\Omega) \to \mathbb{R}$ is linear and continuous. We introduce the functionals $F_n : H^1(\Omega) \to (-\infty, \infty]$

$$F_n(u) = \begin{cases} \dfrac{1}{2}\displaystyle\int_\Omega |\nabla u|^2 \mathrm{d}x + n \int_{\partial\Omega} u^2 \mathrm{d}s - f(u) & \text{for } u \in A_n, \\ \infty & \text{otherwise ,} \end{cases}$$

as well as the Dirichlet energy $F : H^1(\Omega) \to (-\infty, \infty]$

$$F(u) = \begin{cases} \dfrac{1}{2}\displaystyle\int_\Omega |\nabla u|^2 \mathrm{d}x - f(u) & \text{for } u \in H_0^1(\Omega), \\ \infty & \text{otherwise .} \end{cases}$$

Theorem 5 (Γ-convergence)**.** *The sequence $(F_n)_{n\in\mathbb{N}}$ of functionals Γ-converges to the Dirichlet energy F in the weak topology of $H^1(\Omega)$.*

Proof. We start by checking the liminf inequality. Let first $u \notin H_0^1(\Omega)$ and $(u_n)_{n\in\mathbb{N}}$ be any sequence in $H^1(\Omega)$ such that $u_n \rightharpoonup u$. By linearity and continuity of the trace operator it follows that $\operatorname{tr}(u_n) \rightharpoonup \operatorname{tr}(u)$ in $L^2(\partial\Omega)$ and since u has nontrivial boundary values we have $\operatorname{tr}(u) \neq 0$. Using the weak semicontinuity of $\|\cdot\|^2_{L^2(\partial\Omega)}$ it follows that

$$\liminf_{n\to\infty} \|\operatorname{tr}(u_n)\|^2_{L^2(\partial\Omega)} \ge \|\operatorname{tr}(u)\|^2_{L^2(\partial\Omega)} > 0.$$

[3] With this we denote the weak convergence of the sequence $(x_n)_{n\in\mathbb{N}}$ towards x.

Using these facts we establish the liminf inequality in this case

$$\liminf_{n\to\infty} F_n(u_n) \geq \liminf_{n\to\infty} n \left\|\operatorname{tr}(u_n)\right\|^2_{L^2(\partial\Omega)} - \lim_{n\to\infty} f(u_n) = \infty = F(u).$$

Now we treat the case $u \in H^1_0(\Omega)$. Then using weak lower semi-continuity we find

$$\liminf_{n\to\infty} F_n(u_n) \geq \liminf_{n\to\infty} \frac{1}{2}\int_\Omega |\nabla u_n|^2 \,\mathrm{d}x - f(u) \geq \frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x - f(u) = F(u).$$

We are left to construct a recovery sequence. Assume that u does not have zero boundary conditions. Then just choose $u_n = u$ for all n. Otherwise assume u lies in $H^1_0(\Omega)$. By Theorem 2 there is a sequence $(u_n)_{n\in\mathbb{N}}$ such that $u_n \in A_n \cap H^1_0(\Omega)$ and $u_n \to u$ strongly in $H^1(\Omega)$ and hence it follows

$$F_n(u_n) = \frac{1}{2}\int_\Omega |\nabla u_n|^2 \,\mathrm{d}x - f(u_n) \to F(u).$$

□

Now we discuss the requirements of Corollary 4 which will yield the convergence of quasi-minimisers. Firstly, the existence of unique minimisers is well known in the literature of variational problems (see Struwe, 1990). The equi-coercivity of the sequence of functionals $(F_n)_{n\in\mathbb{N}}$ follows from the Poincaré type inequality (4) which can be proved with arguments adapted from Alt (1992). The proofs of both assertions are fully elaborated in the appendix.

Proposition 6. *The sequence $(F_n)_{n\in\mathbb{N}}$ is equi-coercive.*

Lemma 7. *Let $r > 0$ be fixed and consider the set $M \subset H^1(\Omega)$ of functions that satisfy*

$$\frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \int_{\partial\Omega} u^2 \,\mathrm{d}s - f(u) \leq r. \tag{3}$$

Then there exists C only depending on r and Ω such that for every $u \in M$ we have

$$\|u\|_{L^2(\Omega)} \leq C(\|\nabla u\|_{L^2(\Omega)} + 1). \tag{4}$$

An application of Theorem 4 now yields the convergence of quasi minimsers.

Proposition 8. *Any sequence $(u_n)_{n\in\mathbb{N}}$ of quasi-minimizers of $(F_n)_{n\in\mathbb{N}}$ converges to the solution of the Dirichlet problem, both weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$.*

In practice, one will not optimise the functionals F_n over X but rather the objective functions $L_n(\theta) \coloneqq F_n(u_\theta)$ over the parameter space Θ_n. However, if $(\theta_n)_{n\in\mathbb{N}}$ is a sequence of quasi-minimisers of $(L_n)_{n\in\mathbb{N}}$, then $(u_{\theta_n})_{n\in\mathbb{N}}$ is a sequence of quasi-minimisers of $(F_n)_{n\in\mathbb{N}}$. Hence, the previous result yields the convergence of $(u_{\theta_n})_{n\in\mathbb{N}}$ towards the solution of the Dirichlet problem which is our main result.

Theorem 9. *Let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence of quasi-minimisers of $(L_n)_{n\in\mathbb{N}}$. Then $(u_{\theta_n})_{n\in\mathbb{N}}$ converges weakly in $H^1(\Omega)$ and hence strongly in $L^2(\Omega)$ to the solution u of the Dirichlet problem*

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

5 Discussion

We established a convergence result for the deep Ritz method in the following sense. If networks of growing size are trained to quasi-minimise a penalised Dirichlet energy with penalisation strength approaching $+\infty$ these networks converge to the solution of the Dirichlet problem. This result generalises to a wide class of variational energies that include nonlinear PDEs like the p-Laplace operator and to other neural networks that satisfy the universal approximation property.

However, this result does not resolve the study of the deep Ritz method completely since we do not take the numerical evaluation of the Dirichlet energies or the network optimisation into account. Further, it is not clear under what assumptions this approach is able to overcome the curse of dimensionality and whether a similar approach can be taken for higher order stationary PDEs and PDE constraint optimisation problems. Finally, it remains to establish error estimates.

ACKNOWLEDGMENTS

JM acknowledges support by the Evangelisches Studienwerk Villigst e.V. and the IMPRS MiS. The authors want to thank Nihat Ay, Luca Courte and Guido Montúfar for valuable comments.

REFERENCES

H. W. Alt. Linear Functional Analysis. *An application oriented introduction*, 1992.

R. Arora, A. Basu, P. Mianjy, and A. Mukherjee. Understanding Deep Neural Networks with Rectified Linear Units. *arXiv preprint arXiv:1611.01491*, 2016.

J. Berner, P. Grohs, and A. Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *arXiv preprint arXiv:1809.03062*, 2018.

H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer Science & Business Media, 2010.

G. Dal Maso. *An Introduction to* Γ*-Convergence*, volume 8. Springer Science & Business Media, 2012.

W. E and B. Yu. The Deep Ritz Method: A Deep Learning-Based Numerical Algorithm for Solving Variational Problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

W. E, J. Han, and A. Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

M. T. McCann, K. H. Jin, and M. Unser. Convolutional Neural Networks for Inverse Problems in Imaging: A Review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602*, 2013.

P. Petersen, M. Raslan, and F. Voigtlaender. Topological properties of the set of functions generated by neural networks of fixed size. *arXiv preprint arXiv:1806.08459*, 2018.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for Activation Functions. *arXiv preprint arXiv:1710.05941*, 2017.

W. Ritz. Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1909(135):1–61, 1909.

M. Struwe. *Variational Methods*, volume 31999. Springer, 1990.

A UNIVERSAL APPROXIMATION IN SOBOLEV TOPOLOGY

In this section of the appendix we prove the universal approximation result which we stated as Theorem 2 in the main text. Our proof uses that every continuous, piecewise linear function can be represented by a neural network with ReLU activation function and then shows how to approximate Sobolev functions with zero boundary conditions by such functions. The precise definition of a piecwise linear function is the following.

Definition 10 (Continuous piecewise linear function)**.** We say a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is *continuous piecewise linear* or shorter *piecewise linear* if there exists a finite set of closed polyhedra whose union is $\mathbb{R}^d$, and f is affine linear over each polyhedron. Note every piecewise linear functions is continuous by definition since the polyhedra are closed and cover the whole space $\mathbb{R}^d$, and affine functions are continuous.

Theorem 11 (Universal expression)**.** *Every ReLU neural network $u_\theta : \mathbb{R}^d \to \mathbb{R}$ is a piecewise linear function. Conversely, every piecewise linear function $f\colon \mathbb{R}^d \to \mathbb{R}$ can be expressed by a ReLU network of depth at most $\lceil \log_2(d+1)\rceil + 1$.*

For the proof of this statement we refer to Arora et al. (2016). We turn now to the approximation capabilities of piecewise linear functions. For an open set $\Omega \subset \mathbb{R}^d$ we denote by $C_c^\infty(\Omega)$ the space of infinitely often differentiable functions with compact support in Ω. Furthermore, for $p \in [1, \infty]$, the Sobolev space $W^{1,p}(\Omega)$ consists of weakly differentiable functions u such that both u and all its weak derivatives are integrable in p-th power. We refer to Brezis (2010) for more details on these function spaces and their associated norms.

Lemma 12. *Let $\varphi \in C_c^\infty(\mathbb{R}^d)$ be a smooth function with compact support. Then for every $\varepsilon > 0$ there is a piecewise linear function s_ε such that for all $p \in [1, \infty]$ it holds*

$$\|s_\varepsilon - \varphi\|_{W^{1,p}(\mathbb{R}^d)} < \varepsilon \quad \textit{and} \quad \operatorname{supp}(s_\varepsilon) \subset \operatorname{supp}(\varphi) + B_\varepsilon(0).$$

Here we set $B_\varepsilon(0)$ to be the ε-ball around zero, i.e., $B_\varepsilon(0) = \{x \in \mathbb{R} \mid |x| < \varepsilon\}$.

Proof. In the following we will denote by $\|\cdot\|_\infty$ the uniform norm on $\mathbb{R}^d$. To show the assertion choose a triangulation $\mathcal{T}$ of $\mathbb{R}^d$ of width $\delta = \delta(\varepsilon) > 0$, consisting of rotations and translations of one non-degenerate simplex K. We choose s_ε to agree with φ on all vertices of elements in $\mathcal{T}$. Since φ is compactly supported it is uniformly continuous and hence it is clear that $\|\varphi - s_\varepsilon\|_\infty < \varepsilon$ if δ is chosen small enough.
To prove convergence of the gradients we show that also $\|\nabla\varphi - \nabla s_\varepsilon\|_\infty < \varepsilon$ which will be shown on one element $K \in \mathcal{T}$ and as the estimate is independent of K is understood to hold on all of $\mathbb{R}^d$. So let $K \in \mathcal{T}$ be given and denote its vertices by $x_1, \dots, x_{n+1}$. We set $v_i = x_{i+1} - x_1$, $i = 1, \dots, d$ to be the vectors spanning K. By the one dimensional mean value theorem we find ξ_i on the line segment joining x_1 and x_i such that

$$\partial_{v_i} s_\varepsilon(v_1) = \partial_{v_i}\varphi(\xi_i).$$

Note that $\partial_{v_i} s_\varepsilon$ is constant on all of K where it is defined. Now for arbitrary $x \in K$ we compute with setting $w = \sum_{i=1}^d \alpha_i v_i$ for $w \in \mathbb{R}^d$ with $|w| \leq 1$. Note that the α_i are bounded uniformly in w, where we use that all elements are the same up to rotations and translations.

$$\begin{aligned}\|\nabla\varphi(x) - \nabla s_\varepsilon(x)\| &= \sup_{|w|\leq 1} \|\nabla\varphi(x)w - \nabla s_\varepsilon(x)w\| \\ &\leq \sup_{|w|\leq 1} \sum_{i=1}^d |\alpha_i| \underbrace{\|\partial_{v_i}\varphi(x) - \partial_{v_i} s_\varepsilon(x)\|}_{=(*)}\end{aligned}$$

where again $(*)$ is uniformly small due to the uniform continuity of $\nabla\varphi$. Noting that the $W^{1,\infty}$ case implies the claim for all $p \in [1, \infty)$ the proof is complete. □

We turn to the proof of Theorem 2 which we state again for the convenience of the reader. Note that this can be generalised to the Sobolev spaces $W_0^{1,p}(\Omega)$.

Theorem 2 (Universal approximation)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be an open set and $u \in H_0^1(\Omega)$. Then for all $\varepsilon > 0$ there exists a function $u_\varepsilon \in H_0^1(\Omega)$ that can be expressed by a ReLU network of depth $\lceil \log_2(d+1) \rceil + 1$ such that*

$$\|u - u_\varepsilon\|_{H^1(\Omega)} < \varepsilon.$$

Proof. Let $u \in H_0^1(\Omega)$ and $\varepsilon > 0$. By the density of $C_c^\infty(\Omega)$ in $H_0^1(\Omega)$ (see Brezis, 2010) we choose a smooth function $\varphi_\varepsilon \in C_c^\infty(\Omega)$ such that $\|u - \varphi_\varepsilon\| \le \varepsilon/2$. Furthermore we use Lemma 12 and choose a piecewise linear function u_ε such that $\|\varphi_\varepsilon - u_\varepsilon\|_{H^1} \le \varepsilon/2$ and such that u_ε has compact support in Ω. This yields

$$\|u - u_\varepsilon\|_{H^1(\Omega)} \le \|u - \varphi_\varepsilon\|_{H^1(\Omega)} + \|\varphi_\varepsilon - u_\varepsilon\|_{H^1(\Omega)} \le \varepsilon$$

and by Theorem 11 we know that u_ε is in fact a realisation of a neural network with depth at most $\lceil \log_2(d+1) \rceil + 1$. ☐

B Poincaré type inequality and equi-coercivity

This section of the appendix provides the proof of the Poincaré type inequality we stated in the main text and shows how this leads to the equi-coercivity of the sequence (F_n) hence provides the proof to the main result which is Theorem 9 in the main text. We begin by establishing that the limit functional F has a unique minimiser as this is required in Theorem 4.

Lemma 13. *The functional $F : H^1(\Omega) \to (-\infty, \infty]$*

$$F(u) = \begin{cases} \dfrac{1}{2}\displaystyle\int_\Omega |\nabla u|^2 \mathrm{d}x - f(u) & u \in H_0^1(\Omega), \\ \infty & u \in H^1(\Omega) \setminus H_0^1(\Omega). \end{cases}$$

has a unique minimiser.

Proof. The existence follows by the direct method of the calculus of variations (see Struwe, 1990, Chapter 1) and the uniqueness by the strict convexity of F. ☐

The following lemma uses a classical compactness argument to establish a Poincaré type inequality, see for example Alt (1992) in Chapter 8 on generalised Poincaré inequalities.

Lemma 7. *Let $\Omega \subseteq \mathbb{R}^d$ be an open and bounded set with Lipschitz boundary. Let further $r > 0$ be fixed and consider the set $M \subset H^1(\Omega)$ defined as*

$$M := \Big\{ u \in H^1(\Omega) \mid \frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \int_{\partial\Omega} u^2 \,ds - f(u) \le r \Big\}. \tag{5}$$

Those functions satisfy a Poincaré type inequality of the form

$$\|u\|_{L^2(\Omega)} \le C(\|\nabla u\|_{L^2(\Omega)} + 1), \tag{6}$$

where C only depends on r and Ω.

Proof. The proof consists of two steps. First we will show that the inequality (5) implies that M cannot contain arbitrarily large, constant functions and second we prove that a failure of the Poincaré inequality (6) means that M contains any large, constant function hence the assertion follows.

Step 1. Let $\xi \in \mathbb{R}$ be a constant function in M. We will show that there is some fixed $C > 0$ depending only on r, $\|f\|_{H^1(\Omega)'}$ and $|\partial\Omega|$ such that

$$|\xi| \le C.$$

Using a scaled version of Young's inequality with $\varepsilon|\Omega|^{1/2} \leq |\partial\Omega|/2$ we compute

$$\begin{aligned} r \geq \int_{\partial\Omega} \xi^2 \, ds - f(\xi) &\geq |\xi|^2 |\partial\Omega| - \|f\|_{H^1(\Omega)'} \|\xi\|_{H^1(\Omega)} \\ &= |\xi|^2|\partial\Omega| - \|f\|_{H^1(\Omega)'} |\Omega|^{1/2}|\xi| \\ &\geq |\xi|^2|\partial\Omega| - C(\varepsilon) \|f\|^2_{H^1(\Omega)'} - \varepsilon|\Omega| \cdot |\xi|^2 \\ &\geq \frac{1}{2}|\partial\Omega| \, |\xi|^2 - C(\varepsilon) \|f\|^2_{H^1(\Omega)'} . \end{aligned}$$

Thus we can solve for $|\xi|$ and find a uniform bound in terms of r, $\|f\|_{H^1(\Omega)'}$ and $|\partial\Omega|$.

Step 2. Now we assume that the inequality fails and will show that this implies that M contains arbitrarily large constant functions. Assume therefore that there is a sequence $(u_k) \subset M$ such that

$$\|\nabla u_k\|_{L^2(\Omega)} + 1 \leq \frac{1}{k} \|u_k\|_{L^2(\Omega)} .$$

This inequality implies that $\|u_k\|_{L^2(\Omega)} \to \infty$ and hence for every large but fixed $R > 0$ we may assume that $\|u_k\|^{-1}_{L^2(\Omega)} R \leq 1$ and set $v_k = u_k(R \|u_k\|^{-1}_{L^2(\Omega)})$. By the star shape of M the v_k are a sequence in M and the above inequality yields upon multiplying

$$\|\nabla v_k\|_{L^2(\Omega)} + \frac{R}{\|u_k\|_{L^2(\Omega)}} \leq \frac{R}{k} \to 0. \tag{7}$$

As $\|v_k\|_{L^2(\Omega)} = R$ and (7) implies a bound on $\|\nabla v_k\|_{L^2(\Omega)}$ we extract a weakly $H^1(\Omega)$ convergent subsequence $v_j \rightharpoonup v$ with limit v in M by the weak closedness of M. Also from (7) we deduce that

$$\nabla v_j \rightharpoonup \nabla v = 0 \quad \text{weakly in } L^2(\Omega)^n$$

and thus there is a constant $\xi \in \mathbb{R}$ such that $v(x) = \xi$ up to a set of measure zero in Ω. To identify ξ we employ the Rellich compactness theorem (see Alt, 1992) which yields that $v_j \to v$ strongly in $L^2(\Omega)$ and together with $\|v_j\|_{L^2(\Omega)} = R$ we conclude

$$R = \|v\|_{L^2(\Omega)} = \|\xi\|_{L^2(\Omega)} = |\Omega|^{1/2}|\xi|$$

and as $R > 0$ was arbitrary this shows that M contains any large, constant function which manifests the desired contradiction. ☐

Note that Rellich's theorem requires some regularity of the boundary of Ω. We assumed that it locally is the graph of a Lipschitz continuous function but the lemma above holds whenever the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact.

Before we turn to the equi-coercivity we define the functional

$$G_n(u) = \frac{1}{2} \int_\Omega |\nabla u|^2 \, \mathrm{d}x + n \int_{\partial\Omega} u^2 \, ds - f(u)$$

hence G_n agrees with F_n where $F_n \neq \infty$. Note that it trivially holds

$$F_n = G_n + \chi_{A_n}$$

where $\chi_{A_n} = 0$ on A_n and $\chi_{A_n} = \infty$ otherwise. We also remark that

$$F_n \geq G_n \geq G_1.$$

This means to show the equi-coercivity of $(F_n)_{n\in\mathbb{N}}$ it suffices to prove the coercity of the single functional G_1.

Proposition 6. *The sequence $(F_n)_{n\in\mathbb{N}}$ is equi-coercive, i.e., for every $r \in \mathbb{R}$ there is a bounded set $K_r \subset H^1(\Omega)$ such that*

$$\bigcup_{n\in\mathbb{N}} \{F_n \leq r\} \subset K_r.$$

Proof. We fix $r \in \mathbb{R}$. As we discussed before the statement of the theorem it is enough to show that $M = \{G_1 \leq r\} \subset K_r$ for some bounded set K_r in $H^1(\Omega)$. The next observation is that upon adding $\|\nabla u\|_{L^2(\Omega)}$ to the inequality (6) we find that

$$\|u\|_{H^1(\Omega)} \leq C(\|\nabla u\|_{L^2(\Omega)} + 1) \tag{8}$$

hence it is enough to provide an $L^2(\Omega)$ bound on the gradients of the functions in M. We employ a scaled version of Young's inequality with fitting $\varepsilon > 0$ and compute using the inequality (8)

$$\begin{aligned} r &\geq \frac{1}{2}\|\nabla u\|^2_{L^2(\Omega)} - \|f\|_{H^1(\Omega)'}\|u\|_{H^1(\Omega)} \\ &\geq \frac{1}{2}\|\nabla u\|^2_{L^2(\Omega)} - C\|f\|_{H^1(\Omega)'}\|\nabla u\|_{L^2(\Omega)} - C\|f\|_{H^1(\Omega)'} \\ &\geq \frac{1}{2}\|\nabla u\|^2_{L^2(\Omega)} - \varepsilon\|\nabla u\|^2_{L^2(\Omega)} - (C(\varepsilon) + C)\|f\|_{H^1(\Omega)'} \\ &\geq \frac{1}{4}\|\nabla u\|^2_{L^2(\Omega)} - (C(\varepsilon) + C)\|f\|_{H^1(\Omega)'}. \end{aligned}$$

Rearranging the inequality we see that $\|\nabla u\|_{L^2(\Omega)} \leq C(r)$ which is sufficient to conclude the proof due to (8). □

Proposition 8. *Any sequence $(u_n)_{n\in\mathbb{N}}$ of quasi-minimizers of $(F_n)_{n\in\mathbb{N}}$ converges to the solution of the Dirichlet problem, both weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$.*

The proof of this result follows from an application of Theorem 4. In practice, one will not optimise the functionals F_n over X but rather the objective functions $L_n(\theta) := F_n(u_\theta)$ over the parameter space Θ_n. However, the convergence of the neural networks arising from training[4] converge towards the solution of the Dirichlet problem which is our main result.

Theorem 9. *Let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence of quasi-minimisers of $(L_n)_{n\in\mathbb{N}}$. Then $(u_{\theta_n})_{n\in\mathbb{N}}$ converges weakly in $H^1(\Omega)$ and hence strongly in $L^2(\Omega)$ to the solution u of the Dirichlet problem*

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Proof. Let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence of quasi-minimisers of $(L_n)_{n\in\mathbb{N}}$. Since $F_n \equiv \infty$ on the complement of $A_n = \{u_\theta \mid \theta \in \Theta_n\}$ we have

$$\inf_{u\in H^1(\Omega)} F_n(u) = \inf_{\theta\in\Theta_n} F_n(u_\theta) = \inf_{\theta\in\Theta_n} L_n(\theta).$$

Now the computation

$$F_n(u_{\theta_n}) = L_n(\theta_n) \leq \inf_{\theta\in\Theta_n} L_n(\theta) + \delta_n = \inf_{u\in H^1(\Omega)} F_n(u) + \delta_n$$

shows that $(u_{\theta_n})_{n\mathbb{N}}$ is a sequence of quasi-minimisers of $(F_n)_{n\in\mathbb{N}}$. Proposition 8 now yields the claim. □

Note that the assumption of a Lipschitz boundary was only used in the proof of the Poincaré type inequality in Lemma 7. However, this lemma and therefore our main result hold whenever the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact.

C EXTENSION TO GENERAL VARIATIONAL PROBLEMS

We conclude the appendix by briefly considering how to extend our results, in fact the above considerations admit a direct extension to a considerably broader class of energies, including energies associated to higer order elliptic equations and also non-quadratic ones such as the p-Dirichlet energy, the Euler-Lagrange equations of which is the p-Laplace (see Struwe, 1990). In this section the activation function $\rho : \mathbb{R} \to \mathbb{R}$ is no longer assumed to be ReLU but needs to be chosen according to the energy.

[4]or more precisely the functions arising from the parameters obtained through training

Setting 14. We begin with the assumptions and the norm structure on the space where the energy will live. Assume X is a reflexive Banach space with norm $\|\cdot\|_X$ and that there is an additional norm $|\cdot|$ on X, which does not need to render X complete. Furthermore let Y be another Banach space with norm $\|\cdot\|_Y$ and let $T : (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ be a linear and continuous operator which is linked to the norm $\|\cdot\|_X$ in the following way

$$\|x\|_X = |x| + \|Tx\|_Y \quad \text{for all } x \in X.$$

As for the activation function the only thing we require for the moment is that the neural networks are members of the space X and $A_n \subset X$ denotes again the networks of some fixed architecture that is growing in n. Of course this implies that our Banach space X is some kind of function space. We turn to the energy and some abstract analogue of boundary values. We assume that there are two maps

$$E : X \to \mathbb{R} \quad \text{and} \quad \gamma : X \to B$$

where E (the energy) is bounded from below, weakly lower semi-continuous and norm-continuous. Both, the weak topology related to the weak lower semi-continuity and the norm-continuity are meant with respect to the norm $\|\cdot\|_X$. Furthermore γ (the trace operator) is a linear and continuous map from $(X, \|\cdot\|)$ into the Banach space B that is the abstract analogue of boundary values. We set $X_0 = \ker(\gamma)$. With this terminology fixed we are able to define our functionals $F_n, F : X \to (-\infty, \infty]$. Let $f \in X'$ and set

$$F_n(x) = \begin{cases} E(x) + n\,\|\gamma(x)\|_B^2 - f(x) & x \in A_n, \\ \infty & x \notin A_n. \end{cases}$$

and for the limit functional

$$F(x) = \begin{cases} E(x) - f(x) & x \in X_0, \\ \infty & x \notin X_0. \end{cases}$$

To get an intuition for the setting think of $X = H^1(\Omega)$ where $|\cdot| = \|\cdot\|_{L^2}$ and $\|\cdot\|_X = \|\cdot\|_{H^1}$ such as $T : H^1(\Omega) \to L^2(\Omega)^d$ with $u \mapsto \nabla u$. The question we ask now is:

Under which assumptions do we obtain the Γ-convergence and equi-coercivity of $(F_n)_{n\in\mathbb{N}}$ to F in the weak topology of X?

To this end we formulate the hypothesis below.

- (H1) The union $\bigcup_{n\in\mathbb{N}} A_n \cap X_0$ is dense in X_0 with respect to the norm $\|\cdot\|_X$.
- (H2) The space X is reflexive, its norm is given as $\|\cdot\|_X = |\cdot| + \|T\cdot\|_Y$, for some norm $|\cdot|$ on X and $T : (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ is linear and continuous into the Banach space Y.
- (H3) The identity $(X, \|\cdot\|_X) \to (X, |\cdot|)$ maps weakly convergent sequences to strongly convergent ones.
- (H4) The map γ is linear and continuous and the set $\{\|\gamma(x)\|_B^2 - f(x) \le r\} \cap \ker(T)$ is bounded in X by a constant $C = C(r) < \infty$ for every $r \in \mathbb{R}$.
- (H5) The energy $E : X \to \mathbb{R}$ is bounded from below, weakly lower semi-continuous (with respect to the weak topology induced by $\|\cdot\|_X$) and also $\|\cdot\|_X$-continuous and satisfies $|E(x)| \ge c_1 \|Tx\|_Y^p - c_2$ for $p > 1$ and constants c_1, c_2. Furthermore assume that E has a unique minimiser on X_0.

We can again formulate our main result, its proof is very similar to the case of the Dirichlet energy.

Theorem 15. *Under the hypotheses* (H1)-(H5) *the sequence $(F_n)_{n\in\mathbb{N}}$ of functionals Γ-converges towards F. Further every sequence of quasi-minimisers of $(F_n)_{n\in\mathbb{N}}$ converges to the unique minimiser of F.*

Proof. The proof mainly consists of abstraction of the concepts we already met in the Dirichlet case and therefore we will keep it brief. A look back to the shows that (H1), together with the strong continuity of E from (H5) provides a recovery sequence and the $\liminf$ inequality follows by the lower semi-continuity assumed in (H5). The hypothesis (H2)-(H4) take care of the equi-coercivity which deserves some extra comments and we refer the reader to the next lemma. □

Lemma 16 (Abstract Poincaré Inequality)**.** *Consider the setting described above i.e., let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces with a linear and continuous map $T : X \to Y$ and another norm $|\cdot|$ on X such that $\|x\|_X = |x| + \|Tx\|_Y$ and the identity $(X, \|\cdot\|_X) \to (X, |\cdot|)$ maps weakly to strongly convergent sequences. Let M be some weakly closed, star-shaped set with center zero. Then $\ker(T) \cap M$ is bounded if and only if there is a constant C such that*

$$|x| \leq C(\|Tx\|_Y + 1) \quad \text{for all } x \in M.$$

Proof. This works exactly as in the case $X = H^1(\Omega)$ and $Y = L^2(\Omega)^d$ with $T = \nabla$ which we studied before. □

THE p-LAPLACE: AN EXAMPLE FOR A NONLINEAR PDE

We illustrate the abstract setting by considering the p-Dirichlet energy for $p \in (1, \infty)$ given by

$$E : W^{1,p}(\Omega) \to \mathbb{R} \quad \text{with} \quad E(u) = \frac{1}{p} \int_\Omega |\nabla u|^p \, dx$$

Note that for $p \neq 2$ the associated Euler-Lagrange equation is nonlinear. This PDE is called the p-Laplace equation and, in strong formulation, is given by

$$\begin{aligned} -\operatorname{div}(|\nabla u|^{p-2} \nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

compare also to the first chapter of Struwe (1990). Choosing the ReLU activation function, the abstract setting presented in (14) is applicable in this case by the following choices

$$X = W^{1,p}(\Omega), \quad Y = L^p(\Omega)^d, \quad B = L^p(\partial\Omega), \quad |u| = \|u\|_{L^p(\Omega)}$$

as well as

$$\begin{aligned} \gamma = \operatorname{tr} : W^{1,p}(\Omega) &\to L^p(\partial\Omega) \\ T : W^{1,p}(\Omega) &\to L^p(\Omega)^d \quad \text{with} \quad u \mapsto \nabla u. \end{aligned}$$

Clearly (H1), (H2) and (H5) are fulfilled and (H3) is due to Rellich's embedding theorem. We look at (H4) and need to guarantee that for every $f \in W^{1,p}(\Omega)'$ and $r > 0$ the following set is bounded in $W^{1,p}(\Omega)$

$$\left\{ \|\operatorname{tr}(u)\|^2_{L^p(\partial\Omega)} - f(u) \leq r \right\}.$$

This works similar to the case $p = 2$ discussed above and uses again a scaled version of Young's inequality.