Neural Network Approximations of PDEs Beyond Linearity: A Representational Perspective

Tanya Marwah [1] **Zachary C. Lipton** [1] **Jianfeng Lu** [2] **Andrej Risteski** [1]

Abstract

A burgeoning line of research leverages deep neural networks to approximate the solutions to high dimensional PDEs, opening lines of theoretical inquiry focused on explaining how it is that these models appear to evade the curse of dimensionality. However, most prior theoretical analyses have been limited to linear PDEs. In this work, we take a step towards studying the representational power of neural networks for approximating solutions to nonlinear PDEs. We focus on a class of PDEs known as *nonlinear elliptic variational PDEs*, whose solutions minimize an *Euler-Lagrange* energy functional $\mathcal{E}(u) = \int_\Omega L(x, u(x), \nabla u(x)) - f(x)u(x)dx$. We show that if composing a function with Barron norm b with partial derivatives of L produces a function of Barron norm at most $B_L b^p$, the solution to the PDE can be ϵ-approximated in the L^2 sense by a function with Barron norm $O\left((dB_L)^{\max\{p\log(1/\epsilon), p^{\log(1/\epsilon)}\}}\right)$. By a classical result due to (Barron, 1993), this correspondingly bounds the size of a 2-layer neural network needed to approximate the solution. Treating p, ϵ, B_L as constants, this quantity is polynomial in dimension, thus showing neural networks can evade the curse of dimensionality. Our proof technique involves neurally simulating (preconditioned) gradient in an appropriate Hilbert space, which converges exponentially fast to the solution of the PDE, and such that we can bound the increase of the Barron norm at each iterate. Our results subsume and substantially generalize analogous prior results for linear elliptic PDEs over a unit hypercube.

[1]Carnegie Mellon University [2]Duke University. Correspondence to: Tanya Marwah <tmarwah@andrew.cmu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1. Introduction

Scientific applications have become one of the new frontiers for the application of deep learning (Jumper et al., 2021; Tunyasuvunakool et al., 2021; Sønderby et al., 2020). PDEs are a fundamental modeling techniques, and designing neural networks-aided solvers, particularly in high-dimensions, is of widespread usage in many scientific domains (Hsieh et al., 2019; Brandstetter et al., 2022). One of the most common approaches for applying neural networks to solve PDEs is to parametrize the solution as a neural network and minimize a variational objective that represents the solution (Sirignano & Spiliopoulos, 2018; E & Yu, 2017). The hope in doing so is to have a method which computationally avoids the "curse of dimensionality"—i.e., that scales less than exponentially with the ambient dimension.

To date, neither theoretical analysis nor empirical applications have yielded a precise characterization of the range of PDEs for which neural networks-aided methods outperform classical methods. Active research on the empirical side (Han et al., 2018; E et al., 2017; Li et al., 2020a;b) has explored several families of PDEs, e.g., Hamilton-Bellman-Jacobi and Black-Scholes, where neural networks have been demonstrated to outperform classical grid-based methods. On the theory side, a recent line of works (Marwah et al., 2021; Chen et al., 2021; 2022) has considered the following fundamental question:

For what families of PDEs, can the solution be represented by a small neural network?

The motivation for this question is computational: fitting the neural network (by minimizing some objective) is at least as expensive as the neural network required to represent it. Specifically, these works focus on understanding when the approximating neural network can be sub-exponential in size, thus avoiding the curse of dimensionality. However, to date, these results have only been applicable to *linear* PDEs.

In this paper, we take the first step beyond such work, considering a *nonlinear* family of PDEs and study *nonlinear variational PDEs*. These equations have the form $-\mathrm{div}_x(\partial_{\nabla u} L(x, u, \nabla u)) + \partial_u L(x, u, \nabla u) = f$ and are a (very general) family of *nonlinear Euler-Lagrange* equations. Equivalently, the solution to the

PDE is the minimizer of the energy functional $\mathcal{E}(u) = \int_\Omega (L(x, u(X), \nabla u(x)) - f(x)u(x))\, dx$. This paradigm is very general: it originated with Lagrangian formulations of classical mechanics, and for different L, a variety of variational problems can be modeled or learned (Schmidt & Lipson, 2009; Cranmer et al., 2020). These PDEs have a variety of applications in scientific domains, e.g., (non-Newtonian) fluid dynamics (Koleva & Vulkov, 2018), meteorology (Weller et al., 2016), and nonlinear diffusion equations (Burgers, 2013).

Our main result is to show that *when the function L has "low complexity", so does the solution*. The notion of complexity we work with is the *Barron norm* of the function, similar to Chen et al. (2021); Lee et al. (2017). This is a frequently used notion of complexity, as a function with small Barron norm can be represented by a small, two-layer neural network, due to a classical result (Barron, 1993). Mathematically, our proof techniques are based on "neurally unfolding" an iterative preconditioned gradient descent in an appropriate function space: namely, we show that each of the iterates can be represented by a neural network with Barron norm not much worse than the Barron norm of the previous iterate—along with showing a bound on the number of required steps.

Importantly, our results go beyond the typical non-parametric bounds on the size of an approximator network that can be easily shown by classical regularity results of the solution to the nonlinear variational PDEs (De Giorgi, 1957; Nash, 1957; 1958) along with universal approximation results (Yarotsky, 2017).

2. Overview of Results

Let $\Omega := [0, 1]^d$ be a d-dimensional hypercube and let $\partial\Omega$ denote its boundary.

We first define the energy functional whose minimizers are represented by a nonlinear variational PDE—i.e., the Euler-Lagrange equation of the energy functional.

Definition 1 (Energy functional)**.** *For all $u : \Omega \to \mathbb{R}$ such that $u|_{\partial\Omega} = 0$, we consider an energy functional of the following form:*

$$\mathcal{E}(u) = \int_\Omega \Big(L(x, u(x), \nabla u(x)) - f(x)u(x) \Big) dx, \quad (1)$$

where $L : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ and there exist constants $0 < \lambda \le \Lambda$ such that for every $x \in \Omega$ the function $L(x, \cdot, \cdot) : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is smooth and convex, i.e.,

$$\mathrm{diag}([0, \lambda \mathbf{1}_d]) \le \nabla^2_{(y,z)} L(x, y, z) \le \mathrm{diag}([\Lambda, \Lambda \mathbf{1}_d]) \quad (2)$$

for all $(y, z) \in \mathbb{R} \times \mathbb{R}^d$.

Further, we assume that the function $f : \Omega \to \mathbb{R}$ is such that $\|f\|_{L^2(\Omega)} < \infty$. Note that without loss of generality[1] *we assume that $\lambda \le 1/C_p$ (where C_p is the Poincare constant defined in Theorem 2).*

The minimizer $u^\star$ of the energy functional $\mathcal{E}$ exists and is unique. The proof of existence and uniqueness is standard (following essentially along the same lines as Theorem 3.3 in Fernández-Real & Ros-Oton (2020)), and is stated in the following Lemma (with the full proof provided in Section D.1 of the Appendix for completeness).

Lemma 1. *Let $L : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ be the function as defined in Definition 1. Then the minimizer of the energy functional $\mathcal{E}$ exists and is unique.*

Writing down the condition for stationarity, we can derive a (nonlinear) elliptic PDE for the minimizer of the energy functional in Definition 1 .

Lemma 2. *Let $u^\star : \Omega \to \mathbb{R}$ be the unique minimizer for the energy functional in Definition 1. Then for all $\varphi \in H^1_0(\Omega)$, $u^\star$ satisfies the following condition:*

$$\begin{aligned} &D\mathcal{E}[u](\varphi) \\ &= \int_\Omega (\partial_{\nabla u} L(x, u, \nabla u)\nabla\varphi + \partial_u L(x, u, \nabla u)\varphi - f\varphi)\, dx \\ &= 0, \end{aligned} \quad (3)$$

where $d\mathcal{E}[u](\varphi)$ denotes the directional derivative of the energy functional calculated at u in the direction of φ. Thus, the minimizers of the energy functional satisfy the following PDE with Dirichlet boundary condition:

$$\begin{aligned} &D\mathcal{E}(u) \\ &\quad := -\mathrm{div}_{\mathrm{x}}(\partial_{\nabla u} L(x, u, \nabla u)) + \partial_u L(x, u, \nabla u) = f \end{aligned} \quad (4)$$

for all $x \in \Omega$ and $u(x) = 0, \forall x \in \partial\Omega$. Here $\mathrm{div}_{\mathrm{x}}$ denotes the divergence operator.

The proof for the Lemma can be found in Appendix D.2. Here $-\mathrm{div}_{\mathrm{x}}(\partial_{\nabla u} L(\nabla\cdot))$ and $\partial_u L(x, \cdot, \nabla\cdot)$ are operators that acts on a function (in this case u).[2]

Our goal is to determine if the solution to the PDE in (4) can be expressed by a neural network with a small number of parameters. In order do so, we rely on the concept of a *Barron norm*, which measures the complexity of a function in terms of its Fourier representation. We show that if composing with the function partial derivatives of the function L increases the Barron norm of u in a bounded fashion, then the

[1] Since λ is a lower bound on the strong convexity constant. If we choose a weaker lower bound, we can always ensure $\lambda \le 1/C_p$.

[2] For a vector valued function $F : \mathbb{R}^d \to \mathbb{R}^d$ we will denote the divergence operator either by $\mathrm{div}_{\mathrm{x}} F$ or by $\nabla \cdot F$, where $\mathrm{div}_{\mathrm{x}} F = \nabla \cdot F = \sum_{i=1}^d \frac{\partial_i F}{\partial x_i}$

solution to the PDE in (4) will have a bounded Barron norm. The motivation for using this norm is a seminal paper (Barron, 1993), which established that any function with Barron norm C can be ϵ-approximated by a two-layer neural network in the L^2 sense by a 2-layer neural network with size $O(C^2/\epsilon)$, thus evading the curse of dimensionality if C is substantially smaller than exponential in d. Informally, we will show the following result:

Theorem 1 (Informal)**.** *Given the function L in Definition 1, such that composing a function with Barron norm b with $\partial_{\nabla u} L$ or $\partial_u L$ produces a function of Barron norm at most $B_L b^p$ for some constants $B_L, p > 0$. Then, $\forall \epsilon > 0$, the minimizer of the energy functional in Definition 1 can be ϵ-approximated in the L^2 sense by a function with Barron norm*

$$O\left((dB_L)^{\max\{p\log(1/\epsilon), p^{\log(1/\epsilon)}\}}\right).$$

As a consequence, when ϵ, p, B_L are thought of as constants, we can represent the solution to the Euler-Lagrange PDE (4) by a polynomially-sized network, as opposed to an exponentially sized network, which is what we would get by standard universal approximation results and using regularity results for the solutions of the PDE.

We establish this by neurally simulating a preconditioned gradient descent (for a strongly-convex loss) in an appropriate Hilbert space, and show that the Barron norm of each iterate—which is a function—is finite, and at most polynomially bigger than the Barron norm of the previous iterate. We get the final bound by (i) bounding the growth of the Barron norm at every iteration; and (ii) bounding the number of iterations required to reach an ϵ-approximation to the solution. The result in formally stated in Section 5.

3. Related Work

Over the past few years there has been a growing line of work that utilizes neural networks to parameterize the solution to a PDE. Works such as E et al. (2017); E & Yu (2017); Sirignano & Spiliopoulos (2018); Raissi et al. (2017) achieved impressive results on a variety of different applications and have demonstrated the empirical efficacy of neural networks in solving high dimensional PDEs. This is a great and promising direction for solving high dimensional PDEs since erstwhile dominant numerical approaches like the finite differences and finite element methods (LeVeque, 2007) depend primarily upon discretizing the input space, hence limiting their use for problems on low dimensional input space.

Several recent works look into the theoretical analysis into their representational capabilities has also gained a lot of attention. Khoo et al. (2021) show the existence of a network by discretizing the input space into a mesh and then using convolutional NNs, where the size of the layers is exponential in the input dimension. Sirignano & Spiliopoulos (2018) provide a universal approximation result, showing that for sufficiently regularized PDEs, there exists a multilayer network that approximates its solution. (Jentzen et al., 2018; Grohs & Herrmann, 2020; Hutzenthaler et al., 2020) show that provided a better-than-exponential dependence on the input dimension for some specific parabolic PDEs, based on a stochastic representation using the Feynman-Kac Lemma, thus limiting the applicability of their approach to PDEs that have such a probabilistic interpretation.

These representational results can be further be utilized towards analyzing the generalization properties of neural network approximations to PDE solutions. For example, Lu et al. (2021) show the generalization analysis for the Deep Ritz method for elliptic equations like the Poisson equation and (Lu & Lu, 2021) extends their analysis to the Schrodinger eigenvalue problem. Furthermore, Mishra & Molinaro (2020) look at the generalization properties of physics informed neural networks for a linear operators or for non-linear operators with well-defined linearization.

Closest to our work is a recent line of study that has focused on families of PDEs for which neural networks evade the curse of dimensionality—i.e. the solution can be approximated by a neural network with a subexponential size. In Marwah et al. (2021) the authors show that for elliptic PDEs whose coefficients are approximable by neural networks with at most N parameters, a neural network exists that ϵ-approximates the solution and has size $O(d^{\log(1/\epsilon)}N)$. Chen et al. (2021) extends this analysis to elliptic PDEs with coefficients with small Barron norm, and shows that if the coefficients have Barron norm bounded by B, an ϵ-approximate solution exists with Barron norm at most $O(d^{\log(1/\epsilon)}B)$. The work by Chen et al. (2022) derives related results for the Schrödinger equation on the whole space.

As mentioned, while most of previous works show key regularity results for neural network approximations of solution to PDEs, most of their analysis is limited to simple *linear* PDEs. The focus of this paper is towards extending these results to a family of PDEs referred to as nonlinear variational PDEs. This particular family of PDEs consists of many famous PDEs such as $p-$Laplacian (on a bounded domain) and is used to model phenomena like non-Newtonian fluid dynamics and nonlinear diffusion processes. The regularity results for these family of PDEs was posed as Hilbert's XIXth problem. We note that there are classical results like De Giorgi (1957) and Nash (1957; 1958) that provide regularity estimates on the solutions of a nonlinear variational PDE of the form in (4). One can easily use these regularity estimates, along with standard universal approximation results (Yarotsky, 2017) to show that the solutions can be approximated arbitrarily well. However, the size of the resulting networks will be exponentially large (i.e. they will

suffer from the curse of dimensionality)—so are of no use for our desired results.

4. Notation and Definition

In this section we introduce some key concepts and notation that will be used throughout the paper. For a vector $x \in \mathbb{R}^d$ we use $\|x\|_2$ to denote its ℓ_2 norm. $C^\infty(\Omega)$ is the set of function $f : \Omega \to \mathbb{R}$ that are infinitely differentiable. For a function $F(x, y, z)$ of multiple variables we use $\nabla_x F(x, y, z)$ and $\partial_x F(x, y, z)$ to denote the (partial) derivative w.r.t the variable x (we drop the subscript if the function takes in only a single variable). Similarly, Δ_x denotes the Laplacian operator where the derivatives are taken w.r.t $x \in \mathbb{R}^d$. With a slight abuse of notation, if a function $L : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ takes functions u and ∇u as input, we will denote the partial derivatives w.r.t second and third set of coordinates as, $\partial_u L(x, u, \nabla u)$ and $\partial_{\nabla u} L(x, u, \nabla u)$, respectively.

We also define some important function spaces and associated key results below.

Definition 2. *For a vector valued function $g : \mathbb{R} \to \mathbb{R}^d$ we define the $L^p(\Omega)$ norm for $p \in [1, \infty)$ as*

$$\|g\|_{L^p(\Omega)} = \left(\int_\Omega \sum_i^d |g_i(x)|^p \, dx \right)^{1/p},$$

For $p = \infty$ we have

$$\|g\|_{L^\infty(\Omega)} = \max_i \|g_i\|_{L^\infty(\Omega)},$$

Definition 3. *For a domain Ω, the space of functions $H_0^1(\Omega)$ is defined as,*

$$H_0^1(\Omega) := \{g : \Omega \to \mathbb{R} : g \in L^2(\Omega), \nabla g \in L^2(\Omega), g|_{\partial\Omega} = 0\}.$$

The corresponding norm for $H_0^1(\Omega)$ is defined as, $\|g\|_{H_0^1(\Omega)} = \|\nabla g\|_{L^2(\Omega)}$.

Finally, we will make use of the Poincaré inequality throughout several of our results.

Theorem 2 (Poincaré inequality, Poincaré (1890))**.** *For any domain $\Theta \subset \mathbb{R}^d$ which is open and bounded, there exists a constant $C_p > 0$ such that for all $u \in H_0^1(\Theta)$*

$$\|u\|_{L^2(\Theta)} \leq C_p \|\nabla u\|_{L^2(\Theta)}.$$

This constant can be very benignly behaved with dimension for many natural domains—even dimension independent. One such example are convex domains (Payne & Weinberger, 1960), for which $C_p \leq \pi^2 \text{diam}(\Omega)$. Furthermore, for $\Omega = [0, 1]^d$, the value of C_p can be explicitly calculated and is equal to $1/\pi^2 d$. This is a simple calculation, but we include it for completeness as the following lemma (proved in Section D.3):

Lemma 3. *For the domain $\Omega := [0, 1]^d$, the Poincare constant is equal to $\frac{1}{\pi^2 d}$.*

4.1. Barron Norms

For a function $f : [0, 1]^d \to \mathbb{R}$ the Fourier transform is defined as,

$$\hat{f}(\omega) = \int_{[0,1]^d} f(x) e^{-i2\pi x^T \omega} dx, \quad \omega \in \mathbb{N}^d, \tag{5}$$

where $\mathbb{N}^d$ is the set of vectors with natural numbers as coordinates. The inverse Fourier transform of a function is defined as,

$$f(x) = \sum_{\omega \in \mathbb{N}^d} e^{i2\pi x^T \omega} \hat{f}(\omega) \tag{6}$$

The Barron norm is an average of the norm of the frequency vector weighted by the Fourier magnitude $|\hat{f}(\omega)|$.

Definition 4 (Spectral Barron Norm, (Barron, 1993))**.** *Let Γ define a set of functions defined over $\Omega := [0, 1]^d$ such that $\hat{f}(\omega)$ and $\omega \hat{f}(\omega)$ are absolutely summable, i.e.,*

$$\Gamma = \left\{ f : \Omega \to \mathbb{R} : \sum_{\omega \in \mathbb{N}^d} |\hat{f}(\omega)| < \infty, \;\&\; \sum_{\omega \in \mathbb{N}^d} \|\omega\|_2 |\hat{f}(\omega)| < \infty \right\}$$

Then we define the spectral Barron norm $\|\cdot\|_{\mathcal{B}(\Omega)}$ as

$$\|f\|_{\mathcal{B}(\Omega)} = \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) |\hat{f}(\omega)|.$$

The Barron norm can be thought of as an L_1 relaxation of requiring sparsity in the Fourier basis—which is intuitively why it confers representational benefits in terms of the size of a neural network required. We refer to Barron (1993) for a more exhaustive list of the Barron norms of some common function classes.

The main theorem from Barron (1993) formalizes this intuition, by bounding the size of a 2-layer network approximating a function with small Barron norm:

Theorem 3 (Theorem 1, Barron (1993))**.** *Let $f \in \Gamma$ such that $\|f\|_{\mathcal{B}(\Omega)} \leq C$ and μ be a probability measure defined over Ω. There exists $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ and $c_i \in \mathbb{R}$ such that $\sum_{i=1}^k |c_i| \leq 2C$, there exists a function $f_k(x) = \sum_{i=1}^k c_i \sigma\left(a_i^T x + b_i\right)$, such that we have,*

$$\int_\Omega \left(f(x) - f_k(x)\right)^2 \mu(dx) \lesssim \frac{C^2}{k}.$$

Here σ denotes a sigmoidal activation function, i.e., $\lim_{x \to \infty} \sigma(x) = 1$ and $\lim_{x \to -\infty} \sigma(x) = 0$.

Note that while Theorem 3 is stated for sigmoidal activations like *sigmoid* and *tanh* (after appropriate rescaling), the results are also valid for ReLU activation functions, since $\text{ReLU}(x) - \text{ReLU}(x-1)$ is in fact sigmoidal. We will also need to work with functions that do not have Fourier coefficients beyond some size (i.e. are band limited), so we introduce the following definition:

Definition 5. *We will define the set Γ_W as the set of functions whose Fourier coefficients vanish outside a bounded ball, that is*

$$\Gamma_W = \{f : \Omega \to \mathbb{R} : \textit{s.t. } f \in \Gamma, \\ \& \ \forall w, \|w\|_\infty \geq W, \hat{f}(w) = 0\}.$$

Finally, as we will work with vector valued functions, we will also define the Barron norm of a vector-valued function as the maximum of the Barron norms of its coordinates:

Definition 6. *For a vector valued function $g : \Omega \to \mathbb{R}^d$, we define $\|g\|_{\mathcal{B}(\Omega)} = \max_i \|g_i\|_{\mathcal{B}(\Omega)}$.*

5. Main Result

Before stating the main result we introduce the key assumption.

Assumption 1. *The function L in Definition 1 can be approximated by a function $\tilde{L} : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ such that there exists a constant $\epsilon_L \in [0, \lambda)$ for all $x \in \Omega$ and $u \in H_0^1(\Omega)$ define $q := (x, u(x), \nabla u(x)) \in \Omega \times \mathbb{R} \times \mathbb{R}^d$*

$$\sup_q \|\partial_u L(q) - \partial_u \tilde{L}(q)\|_2 \leq \epsilon_L \|u(x)\|_2, \\ \textit{and, } \sup_q \|\partial_{\nabla u} L(q) - \partial_{\nabla u} \tilde{L}(q)\|_2 \leq \epsilon_L \|u(x)\|_2,$$

Furthermore, we assume that $\tilde{L}$ is such that for all $g \in H_0^1(\Omega)$, we have $\tilde{L}(x, g, \nabla g) \in H_0^1(\Omega)$, $\tilde{L}(x, g, \nabla g) \in \Gamma$ and for all $x \in \Omega$

$$\|\partial_u \tilde{L}(x, g, \nabla g)\|_{\mathcal{B}(\Omega)} \leq B_{\tilde{L}} \|g\|_{\mathcal{B}(\Omega)}^{p_{\tilde{L}}}, \\ \textit{and, } \|\partial_{\nabla u} \tilde{L}(x, g, \nabla g)\|_{\mathcal{B}(\Omega)} \leq B_{\tilde{L}} \|g\|_{\mathcal{B}(\Omega)}^{p_{\tilde{L}}}. \tag{7}$$

for some constants $B_{\tilde{L}} \geq 0$, and $p_{\tilde{L}} \geq 0$. Finally, if $g \in \Gamma_W$ then $\partial_u \tilde{L}(x, g, \nabla g) \in \Gamma_{k_{\tilde{L}} W}$ and $\partial_{\nabla u} \tilde{L}(x, g, \nabla g) \in \Gamma_{k_{\tilde{L}} W}$ for a $k_{\tilde{L}} > 0$.

We refer to *Remark 4* for an example of how the conditions in the assumption manifest for a linear elliptic PDE.

This assumption is fairly natural: it states that the function L is such that its partial derivatives w.r.t u and ∇u can be approximated (up to ϵ_L) by a function $\tilde{L}$ with partial derivatives that have the property that when applied to a function g with small Barron norm, the new Barron norm is not much bigger than that of g. The constant p specifies the order of this growth. The functions for which our results are most interesting are when the dependence of $B_{\tilde{L}}$ on d is at most polynomial—so that the final size of the approximating network does not exhibit curse of dimensionality. For instance, we can take L to be a multivariate polynomial of degree up to P: we show in Lemma 10 the constant $B_{\tilde{L}}$ is $O(d^P)$ (intuitively, this dependence comes from the total number of monomials of this degree), whereas p and k are both $O(P)$.

With all the assumptions stated, we now state our main theorem,

Theorem 4 (Main Result)**.** *Consider the nonlinear variational PDE in* (4) *which satisfies Assumption 1 and let $u^\star \in H_0^1(\Omega)$ denote the unique solution to the PDE. If $u_0 \in H_0^1(\Omega)$ is a function such that $u_0 \in \Gamma_{W_0}$, then for all sufficiently small $\epsilon > 0$, and*

$$T := \left\lceil \log\left(\frac{2}{\epsilon}\frac{\mathcal{E}(u_0) - \mathcal{E}(u^\star)}{\lambda}\right) / \log\left(\frac{1}{1 - \frac{\lambda^6}{(1+C_p)^{10}\Lambda^5}}\right)\right\rceil,$$

there exists a function $u_T \in H_0^1(\Omega)$ such that $u_T \in \Gamma_{(2\pi k_{\tilde{L}})^T W_0}$ with Barron norm $\|u_T\|_{\mathcal{B}(\Omega)}$ bounded by

$$\left((1 + \eta 2\pi k_{\tilde{L}} W_0 (2\pi k_{\tilde{L}} d + 1) B_{\tilde{L}}) \left(1 + \eta \|f\|_{\mathcal{B}(\Omega)}\right)\right)^{pt + \frac{p^t - 1}{p-1}} \\ \cdot \left(\max\{1, \|u_0\|_{\mathcal{B}(\Omega)}^{p^t}\}\right). \tag{8}$$

Furthermore u_T satisfies $\|u_T - u^\star\|_{H_0^1(\Omega)} \leq \epsilon + \tilde{\epsilon}$ where,

$$\tilde{\epsilon} \leq \frac{\epsilon_L R}{\epsilon_L + \Lambda}\left(\left(1 + \eta(1 + C_p)^2 (\epsilon_L + \Lambda)\right)^T - 1\right),$$

where $R := \|u^\star\|_{H_0^1(\Omega)} + \frac{1}{\lambda}\mathcal{E}(u_0)$ and $\eta = \frac{\lambda^4}{4(1+C_p)^7\Lambda^4}$.

Remark 1: The function u_0 can be seen as an initial estimate of the solution, that can be refined to an estimate u_T, which is progressively better at the expense of a larger Barron norm. A trivial choice could be $u_0 = 0$, which has Barron norm 1, and which by Lemma 4 would result in $\mathcal{E}(u_0) \leq \Lambda \|u^*\|_{H_0^1(\Omega)}^2$.

Remark 2: The final approximation error has two terms, and note that T goes to infinity as ϵ tends to zero and is a consequence of the way u_T is constructed — by simulating a functional (preconditioned) gradient descent which converges to the solution to the PDE. $\tilde{\epsilon}$ stems from the approximation that we make between $\tilde{L}$ and L, which grows as T increases — it is a consequence of the fact that the gradient descent updates with $\tilde{L}$ and L progressively drift apart as $T \to \infty$.

Remark 3: As in the informal theorem, if we think of $p, \Lambda, \lambda, C_p, k, \|u_0\|_{\mathcal{B}(\Omega)}$ as constants, the theorem implies that $u^\star$ can be ϵ-approximated in the L^2 sense by a function with Barron norm $O\left((dB_L)^{\max\{p\log(1/\epsilon), p^{\log(1/\epsilon)}\}}\right)$.

Therefore, combining results from Theorem 4 and Theorem 3 the total number of parameters required to $\epsilon-$approximate the solution $u^\star$ by a $2-$layer neural network is

$$O\left(\frac{1}{\epsilon^2}\left(dB_L\right)^{2\max\{p\log(1/\epsilon),p^{\log(1/\epsilon)}\}}\right).$$

Remark 4: The theorem recovers (and vastly generalizes) prior results which bound the Barron norm of linear elliptic PDEs like Chen et al. (2021) over the hypercube. In these results, the elliptic PDE takes the form that for all $u \in H_0^1(\Omega)$, $-\mathrm{div_x}(A\nabla u) + cu = f$ and the functions $A : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $c : \mathbb{R}^d \to \mathbb{R}$ are such that $\forall x \in \Omega$, $A(x)$ is positive definite and $c(x)$ is non-negative and bounded. Further, the functions A and c are assumed to have bounded Barron norm. To recover this setting from our result, consider choosing

$$L(x, u(x), \nabla u(x)) := \frac{1}{2}(\nabla u(x))^T A(x)(\nabla u(x)) + \frac{1}{2}c(x)u(x)^2.$$

For this L, we have $\partial^2_{\nabla u} L(x, u(x), \nabla u(x)) = A(x)$ and $\partial^2_u L(x, u(x), \nabla u(x)) = c(x)$. The conditions in Equation 2 in Definition 1 require that $\lambda \le A(x) \le \Lambda$ and $0 \le c(x) \le \Lambda$, which match the conditions on the coefficients A and c in Chen et al. (2021).

Further, by a simple application of Lemma 8, one can show, $\|\partial_{\nabla u} L(x, u, \nabla u)\|_{\mathcal{B}(\Omega)} \le d^2\|A\|_{\mathcal{B}(\Omega)}\|u\|_{\mathcal{B}(\Omega)}$, and $\|\partial_u L(x, u, \nabla u)\|_{\mathcal{B}(\Omega)} \le \|A\|_{\mathcal{B}(\Omega)}\|u\|_{\mathcal{B}(\Omega)}$ and therefore satisfy (7) in Assumption 1 with $B_{\tilde{L}} = \max\{d^2\|A\|_{\mathcal{B}(\Omega)}, \|c\|_{\mathcal{B}(\Omega)}\}$ and $p = 1$. Plugging these quantities in Theorem 4, we recover the exact same bound from Chen et al. (2021).

6. Proof of Main Result

The proof will proceed by "neurally unfolding" a preconditioned gradient descent on the objective $\mathcal{E}$ in the Hilbert space $H_0^1(\Omega)$. This is inspired by previous works by Marwah et al. (2021); Chen et al. (2021) where the authors show that for a linear elliptic PDE, an objective which is quadratic can be designed. In our case, we show that $\mathcal{E}$ is "strongly convex" in some suitable sense — thus again, bounding the amount of steps needed.

More precisely, the result will proceed in two parts:

1. First, we will show that the sequence of functions $\{u_t\}_{t=0}^\infty$, where $u_{t+1} \leftarrow u_t - \eta(I-\Delta_x)^{-1}d\mathcal{E}(u_t)$ can be interpreted as performing preconditioned gradient descent, with the (constant) preconditioner $(I-\Delta_x)^{-1}$. We show that in some appropriate sense (Lemma 4), $\mathcal{E}$ is strongly convex in $H_0^1(\Omega)$ — thus the updates converge at a rate of $O(\log(1/\epsilon))$.

2. We then show that the Barron norm of each iterate u_{t+1} can be bounded in terms of the Barron norm of the prior iterate u_t. We show this in Lemma 7, where we show that given Assumption 1, $\|u_{t+1}\|_{\mathcal{B}(\Omega)}$ can be bounded as $O(d\|u_t\|^p_{\mathcal{B}(\Omega)})$. By unrolling this recursion we show that the Barron norm of the ϵ-approximation of $u^\star$ is of the order $O(d^{p^T}\|u_0\|^p_{\mathcal{B}(\Omega)})$ where T are the total steps required for ϵ-approximation and $\|u_0\|_{\mathcal{B}(\Omega)}$ is the Barron norm of the first function in the iterative updates.

We now proceed to delineate the main technical ingredients for both of these parts.

6.1. Convergence Rate of Sequence

The proof to show the convergence to the solution $u^\star$ is based on adapting the standard proof (in finite dimension) for convergence of gradient descent when minimizing a strongly convex function f. Recall, the basic idea is to Taylor expand $f(x+\delta) \approx f(x) + \nabla f(x)^T\delta + O(\|\delta\|^2)$. Taking $\delta = \eta\nabla f(x)$, we lower bound the progress term $\eta\|\nabla f(x)\|^2$ using the convexity of f, and upper bound the second-order term $\eta^2\|\nabla f(x)\|^2$ using the smoothness of f.

We follow analogous steps, and prove that we can lower bound the progress term by using some appropriate sense of convexity of $\mathcal{E}$, and upper bound using some appropriate sense of smoothness of $\mathcal{E}$, when considered as a function over $H_0^1(\Omega)$. Precisely, we show:

Lemma 4 (Strong convexity of $\mathcal{E}$ in H_0^1)**.** *If $\mathcal{E}, L$ are as in Definition 1, we have*

1. $\forall u, v \in H_0^1(\Omega) : \langle D\mathcal{E}(u), v\rangle_{L^2(\Omega)} = \int_\Omega \left(-\mathrm{div_x}(\partial_{\nabla u}L(x, u, \nabla u)) + \partial_u(x, u, \nabla u)\right) v dx = \int_\Omega \partial_{\nabla u}L(x, u, \nabla u)\cdot\nabla v + \partial_u L(x, u, \nabla u)v\, dx.$

2. $\forall u, v \in H_0^1(\Omega) : \lambda\|u-v\|^2_{H_0^1(\Omega)} \le \langle D\mathcal{E}(u) - D\mathcal{E}(v), u-v\rangle_{L^2(\Omega)} \le (1+C_p^2)\Lambda\|u-v\|^2_{H_0^1(\Omega)}.$

3. $\forall u, v \in H_0^1(\Omega) : \frac{\lambda}{2}\|\nabla v\|^2_{L^2(\Omega)} + \langle D\mathcal{E}(u) - f, v\rangle_{L^2(\Omega)} \le \mathcal{E}(u+v) - \mathcal{E}(u) \le \langle D\mathcal{E}(u) - f, v\rangle_{L^2(\Omega)} + \frac{(1+C_p)^2\Lambda}{2}\|\nabla v\|^2_{L^2(\Omega)}.$

4. $\forall u \in H_0^1(\Omega) : \frac{\lambda}{2}\|u-u^\star\|^2_{H_0^1(\Omega)} \le \mathcal{E}(u) - \mathcal{E}(u^\star) \le \frac{(1+C_p)^2\Lambda}{2}\|u-u^\star\|^2_{H_0^1(\Omega)}.$

Part 1 is a helpful way to rewrite an inner product of a "direction" v with $D\mathcal{E}(u)$—it is essentially a consequence of integration by parts and the Dirichlet boundary condition. Part 2 and 3 are common proxies of convexity and smoothness: they are ways of formalizing the notion that $\mathcal{E}$ is strongly convex has "Lipschitz gradients", when viewed as a function over $H_0^1(\Omega)$. Finally, Part 4 is a consequence of strong

convexity, capturing the fact that if the value of $\mathcal{E}(u)$ is sub-optimal, u must be (quantitatively) far from u^*. The proof of the Lemma can be found in Appendix A.1.

When analyzing gradient descent in (finite dimensions) to minimize a loss function $\mathcal{E}$, the standard condition for progress is that the inner product of the gradient with the direction towards the optimum is lower bounded as $\langle D\mathcal{E}(u), u^* - u\rangle_{L^2(\Omega)} \geq \alpha\|u - u^*\|^2_{L^2(\Omega)}$ (we have $L^2(\Omega)$ inner product vs $H^1_0(\Omega)$ norm). From Parts 2 and 3 of Lemma 4 one can readily see that the above condition is only satisfied "with the wrong norm": i.e. we only have $\langle D\mathcal{E}(u), u^* - u\rangle_{L^2(\Omega)} \geq \alpha\|u - u^*\|^2_{H^1_0(\Omega)}$. Moreover, since in general, $\|\nabla g\|_{L^2(\Omega)}$ can be arbitrarily bigger than $\|g\|_{L^2(\Omega)}$, there is no way to upper bound the $H^1_0(\Omega)$ norm by the $L^2(\Omega)$ norm.

We can fix this mismatch by instead doing preconditioned gradient, using the fixed preconditioner $(I - \Delta_x)^{-1}$. Towards that, the main lemma about the preconditioner we will need is the following one:

Lemma 5 (Norms with preconditioning)**.** *For all $u \in H^1_0(\Omega)$ we have*

1. $\|(I - \Delta_x)^{-1}\nabla_x \cdot \nabla_x u\|_{L^2(\Omega)} = \|(I - \Delta_x)^{-1}\Delta_x u\|_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)}$.
2. $\|(I - \Delta_x)^{-1}u\|_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)}$
3. $\langle (I - \Delta_x)^{-1}u, u\rangle_{L^2(\Omega)} \geq \frac{1}{1+C_p}\langle(-\Delta_x)^{-1}u, u\rangle_{L^2(\Omega)}$.

The first part of the lemma is a relatively simple consequence of the fact that Δ_x and ∇_x "commute", thus can be re-ordered, and the second part that the operator $(I - \Delta_x)^{-1}$ only decreases the $H^1_0(\Omega)$ norm. The latter lemma can be understood intuitively as $(I - \Delta_x)^{-1}$ and Δ_x^{-1} act as similar operators on eigenfunctions of Δ_x with large eigenvalues (the extra I does not do much) – and are only different for eigenfunctions for small eigenvalues. However, since the smallest eigenvalue is lower bounded by $1/C_p$, their gap can be bounded.

Combining Lemma 4 and Lemma 5, we can show that preconditioned gradient descent exponentially converges to the solution to the nonlinear variational PDE in 4.

Lemma 6 (Convergence of Preconditioned Gradient Descent)**.** *Let $u^\star$ denote the unique solution to the PDE in Definition 4 For all $t \in \mathbb{N}$, we define the sequence of functions*

$$u_{t+1} \leftarrow u_t - \eta(I - \Delta_x)^{-1}\left(D\mathcal{E}(u_t) - f\right). \tag{9}$$

where $\eta = \frac{\lambda^4}{4(1+C_p)^7\Lambda^4}$. If $u_0 \in H^1_0(\Omega)$, then after t iterations we have,

$$\mathcal{E}(u_{t+1}) - \mathcal{E}(u^\star) \leq \left(1 - \frac{\lambda^6}{(1+C_p)^{10}\Lambda^5}\right)\left(\mathcal{E}(u_0) - \mathcal{E}(u^\star)\right).$$

The complete proof for convergence can be found in Section A.3 of the Appendix.

Therefore, using the result from Lemma 4 part 4, i.e., $\|u_t - u^\star\|^2_{H^1_0(\Omega)} \leq \frac{2}{\lambda}\left(\mathcal{E}(u_t) - \mathcal{E}(u^\star)\right)$, we have

$$\|u_t - u^\star\|^2_{H^1_0(\Omega)} \leq \frac{2}{\lambda}\left(1 - \frac{\lambda^6}{(1+C_p)^{10}\Lambda^5}\right)^t\left(\mathcal{E}(u_0) - \mathcal{E}(u^\star)\right).$$

and $\|u_T - u^\star\|^2_{H^1_0(\Omega)} \leq \epsilon$ after T steps, where,

$$T \geq \log\left(\frac{\mathcal{E}(u_0) - \mathcal{E}(u^\star)}{\lambda\epsilon/2}\right) / \log\left(\frac{1}{1 - \frac{\lambda^6}{(1+C_p)^{10}\Lambda^5}}\right). \tag{10}$$

6.2. Bounding the Barron Norm

Having obtained a sequence of functions that converge to the solution $u^\star$, we bound the Barron norms of the iterates. We draw inspiration from Marwah et al. (2021); Lu et al. (2021) and show that the Barron norm of each iterate in the sequence increases the Barron norm of the previous iterate in a bounded fashion. Note that in general, the Fourier spectrum of a composition of functions cannot easily be expressed in terms of the Fourier spectrum of the functions being composed. However, from Assumption 1 we know that the function L can be approximated by $\tilde{L}$ such that $\partial_{\nabla u}\tilde{L}(x, u, \nabla u)$ and $\partial_u L(x, u, \nabla u)$ increases the Barron norm of u in a bounded fashion. Thus, if we instead of tracking the iterates in (28) we track

$$\tilde{u}_{t+1} = \tilde{u}_t - \eta\,(I - \Delta)^{-1}\,D\tilde{\mathcal{E}}(\tilde{u}_t). \tag{11}$$

we can derive the following result (the proof is deferred to Section C.1 of the Appendix):

Lemma 7. *For the updates in* (11)*, if $\tilde{u}_t \in \Gamma_{W_t}$ then for all $\eta \in (0, \eta]$ we have $\tilde{u}_{t+1} \in \Gamma_{k_{\tilde{L}}W_t}$ and the Barron norm $\|\tilde{u}_{t+1}\|_{|\mathcal{B}(\Omega)}$ can be bounded as follows,*

$$(1 + \eta(2\pi k_{\tilde{L}}d + 1)B_{\tilde{L}}(2\pi W_t)^{p_{\tilde{L}}})\,\|\tilde{u}\|^{p_{\tilde{L}}}_{\mathcal{B}(\Omega)} + \eta\|f\|_{\mathcal{B}(\Omega)}.$$

The proof consists of using the result in (7) about the Barron norm of composition of a function with $\tilde{L}$, as well as counting the increase in the Barron norm of a function by any basic algebraic operation, as established in Lemma 8. Precisely we show:

Lemma 8 (Barron norm algebra)**.** *If $g, g_1, g_2 \in \Gamma$, then the following set of results hold,*

- *Addition:* $\|g_1 + g_2\|_{\mathcal{B}(\Omega)} \leq \|g_1\|_{\mathcal{B}(\Omega)} + \|g_2\|_{\mathcal{B}(\Omega)}$.
- *Multiplication:* $\|g_1 \cdot g_2\|_{\mathcal{B}(\Omega)} \leq \|g_1\|_{\mathcal{B}(\Omega)}\|g_2\|_{\mathcal{B}(\Omega)}$

- *Derivative: if* $h \in \Gamma_W$ *for* $i \in [d]$ *we have* $\|\partial_i g\|_{\mathcal{B}(\Omega)} \leq 2\pi W \|g\|_{\mathcal{B}(\Omega)}$.
- *Preconditioning: if* $g \in \Gamma$, *then* $\|(I-\Delta)^{-1} g\|_{\mathcal{B}(\Omega)} \leq \|g\|_{\mathcal{B}(\Omega)}$.

The proof for the above lemma can be found in Appendix C.4. It bears similarity to an analogous result in (Chen et al., 2021), with the difference being that our bounds are defined in the *spectral* Barron space which is different from the definition of the Barron norm used in (Chen et al., 2021). Other than preconditioning, the other properties follow by a straightforward calculation. For preconditioning, the main observation is that $(I-\Delta)^{-1}$ acts as a diagonal operator in the Fourier basis—thus the Fourier coefficients of $(I-\Delta)^{-1}h$ can be easily expressed in terms of those of h.

Expanding on the recurrence in Lemma 8 we can bound the Barron norm of the function u_T after T iterations as:

Lemma 9. *Given the updates in* (11) *and function* $u_0 \in \Gamma_{W_0}$ *with Barron norm* $\|u_0\|_{\mathcal{B}(\Omega)}$, *then after* T *iterations we have* $\tilde{u}_T \in \Gamma_{(2\pi k_{\tilde{L}})^T W_0}$ *and* $\|u_0\|_{\mathcal{B}(\Omega)}$ *is bounded by,*

$$\left((1+\eta 2\pi k_{\tilde{L}} W_0 (2\pi k_{\tilde{L}} d+1) B_{\tilde{L}})\left(1+\eta\|f\|_{\mathcal{B}(\Omega)}\right)\right)^{pt+\frac{p^t-1}{p-1}} \cdot \left(\max\{1, \|u_0\|_{\mathcal{B}(\Omega)}^{p^t}\}\right) \tag{12}$$

Finally, we exhibit a natural class of functions that satisfy the main Barron growth property in Equations 7. Precisely, we show (multivariate) polynomials of bounded degree have an effective bound on p and B_L:

Lemma 10. *Let* $f(x) = \sum_{\alpha,|\alpha|\leq P}\left(A_\alpha \prod_{i=1}^d x_i^{\alpha_i}\right)$ *where* α *is a multi-index and* $x \in \mathbb{R}^d$. *If* $g : \mathbb{R}^d \to \mathbb{R}^d$ *is such that* $g \in \Gamma_W$, *then we have* $f \circ g \in \Gamma_{PW}$ *and the Barron norm can be bounded as* $\|f \circ g\|_{\mathcal{B}(\Omega)} \leq d^{P/2}\left(\sum_{\alpha,|\alpha|\leq P}|A_\alpha|^2\right)^{1/2}\|g\|_{\mathcal{B}(\Omega)}^P$

Hence if $\tilde{L}$ is a polynomial of degree P then using the fact that for a functions $g : \Omega \to \mathbb{R}$ such that $g \in \Gamma_W$, from Lemma 8 $\max\{\|g\|_{\mathcal{B}(\Omega)}, \|\nabla g\|_{\mathcal{B}(\Omega)}\} \leq 2\pi W \|g\|_{\mathcal{B}(\Omega)}$, we will have

$$\|\tilde{L}(x, g, \nabla g)\|_{\mathcal{B}(\Omega)} \leq d^{P/2}\left(\sum_{\alpha,|\alpha|\leq P}|A_\alpha|^2\right)^{1/2}(2\pi W)^P \|g\|_{\mathcal{B}(\Omega)}^P.$$

Using the *derivative* result from Lemma 8, the constants in Assumption 1 will take the following values $B_{\tilde{L}} = d^{P/2}(2\pi W)^{P+1}\left(\sum_{\alpha,|\alpha|\leq P}|A_\alpha|^2\right)^{1/2}$, and $r = 2\pi W P$.

Finally, since we are using an approximation of the function L we will incur an error at each step of the iteration. The following Lemma shows that the error between the iterates u_t and the approximate iterates $\tilde{u}_t$ increases with t. The error is calculated by recursively tracking the error between u_t and $\tilde{u}_t$ for each t in terms of the error at $t-1$. Note that this error can be controlled by using smaller values of η.

Lemma 11. *Let* $\tilde{L} : \mathbb{R}^d \to \mathbb{R}$ *be the function satisfying the properties in Assumption 1 and we have*

$$\mathcal{E}(u) = \int_\Omega L(x, u(x), \nabla u(x)) - f(x)u(x)\, dx$$
$$\text{and } \tilde{\mathcal{E}}(u) = \int_\Omega \tilde{L}(x, u(x), \nabla u(x)) - f(x)u(x) dx.$$

For $\eta \in (0, \frac{\lambda^4}{4(1+C_p)^7\Lambda^4}]$ *consider the sequences,*

$$u_{t+1} = u_t - \eta(I-\Delta)^{-1}D\mathcal{E}(u_t),$$
$$\text{and, } \tilde{u}_{t+1} = \tilde{u}_t - \eta(I-\Delta)^{-1}D\tilde{\mathcal{E}}(u_t)$$

then for all $t \in \mathbb{N}$ *and denoting* $R := \|u^\star\|_{H_0^1(\Omega)} + \frac{1}{\lambda}\mathcal{E}(u_0)$ *we have,*

$$\|u_t - \tilde{u}_t\|_{H_0^1(\Omega)} \leq \frac{\epsilon_L R}{\epsilon_L + \Lambda}\left(\left(1+\eta(1+C_p)^2(\epsilon_L+\Lambda)\right)^t - 1\right)$$

7. Conclusion and Future Work

In this work, we take a representational complexity perspective on neural networks, as they are used to approximate solutions of nonlinear elliptic variational PDEs of the form $-\mathrm{div}_x(\partial_{\nabla u}L(x,u,\nabla u)) + \partial_u L(x,u,\nabla u) = f$. We prove that if L is such that composing partial derivatives of L with function of bounded Barron norm increases the Barron norm in a bounded fashion, then we can bound the Barron norm of the solution $u^\star$ to the PDE—potentially evading the curse of dimensionality depending on the rate of this increase. Our results subsume and vastly generalize prior work on the linear case (Marwah et al., 2021; Chen et al., 2021) when the domain is a hypercube. Our proof consists of neurally simulating preconditioned gradient descent on the energy function defining the PDE, which we prove is strongly convex in an appropriate sense.

There are many potential avenues for future work. Our techniques (and prior techniques) strongly rely on the existence of a variational principle characterizing the solution of the PDE. In classical PDE literature, these classes of PDEs are also considered better behaved: e.g. proving regularity bounds is much easier for such PDEs (Fernández-Real & Ros-Oton, 2020). There are many non-linear PDEs that come without a variational formulation for which regularity estimates are derived using non-constructive methods like

comparison principles. It is a wide open question to construct representational bounds for any interesting family of PDEs of this kind. It is also a very interesting question to explore other notions of complexity—e.g. number of parameters in a (potentially deep) network like in (Marwah et al., 2021), Rademacher complexity, among others.

8. Acknowledgements

TM is supported by CMU Software Engineering Institute via Department of Defense under contract FA8702-15-D-0002. ZL is supported in part by by Amazon AI, Salesforce Research, Facebook, UPMC, Abridge, the PwC Center, the Block Center, the Center for Machine Learning and Health, and the CMU Software Engineering Institute (SEI) via Department of Defense contract FA8702-15-D-0002. JL is supported in part by NSF award DMS-2012286, and AR is supported in part by NSF award IIS-2211907, an Amazon Research Award, and the CMU/PwC DT&I Center.

References

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Brandstetter, J., Worrall, D., and Welling, M. Message passing neural PDE solvers. *arXiv preprint arXiv:2202.03376*, 2022.

Burgers, J. M. *The nonlinear diffusion equation: asymptotic solutions and statistical problems*. Springer Science & Business Media, 2013.

Chen, Z., Lu, J., and Lu, Y. On the representation of solutions to elliptic PDEs in Barron spaces. *Advances in Neural Information Processing Systems*, 34, 2021.

Chen, Z., Lu, J., Lu, Y., and Zhou, S. A regularity theory for static Schrödinger equations on $\mathbb{R}^d$ in spectral Barron spaces. *arXiv preprint arXiv:2201.10072*, 2022.

Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

De Giorgi, M. d. E. Sulla differenziabilitae l'analiticita delle estremali degli integrali multipli regolari. *Ennio De Giorgi*, pp. 167, 1957.

E, W. and Yu, B. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *arXiv preprint arXiv:1710.00211*, 2017.

E, W., Han, J., and Jentzen, A. 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.

Evans, L. C. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Fernández-Real, X. and Ros-Oton, X. Regularity theory for elliptic PDE. *Forthcoming book*, 2020.

Grohs, P. and Herrmann, L. Deep neural network approximation for high-dimensional elliptic PDEs with boundary conditions. *arXiv preprint arXiv:2007.05384*, 2020.

Han, J., Jentzen, A., and E, W. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34): 8505–8510, 2018.

Hsieh, J.-T., Zhao, S., Eismann, S., Mirabella, L., and Ermon, S. Learning neural PDE solvers with convergence guarantees. *arXiv preprint arXiv:1906.01200*, 2019.

Hutzenthaler, M., Jentzen, A., Kruse, T., and Nguyen, T. A. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN partial differential equations and applications*, 1(2):1–34, 2020.

Jentzen, A., Salimova, D., and Welti, T. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *arXiv preprint arXiv:1809.07321*, 2018.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Khoo, Y., Lu, J., and Ying, L. Solving parametric PDE problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

Koleva, M. N. and Vulkov, L. G. Numerical solution of the Monge-Ampère equation with an application to fluid dynamics. In *AIP Conference Proceedings*, volume 2048, pp. 030002. AIP Publishing LLC, 2018.

Lee, H., Ge, R., Ma, T., Risteski, A., and Arora, S. On the ability of neural nets to express distributions. In *Conference on Learning Theory*, pp. 1271–1296. PMLR, 2017.

LeVeque, R. J. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020a.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020b.

Lu, J. and Lu, Y. A priori generalization error analysis of two-layer neural networks for solving high dimensional Schrödinger eigenvalue problems. *arXiv preprint arXiv:2105.01228*, 2021.

Lu, J., Lu, Y., and Wang, M. A priori generalization analysis of the deep Ritz method for solving high dimensional elliptic equations. *arXiv preprint arXiv:2101.01708*, 2021.

Marwah, T., Lipton, Z., and Risteski, A. Parametric complexity bounds for approximating PDEs with neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Mishra, S. and Molinaro, R. Estimates on the generalization error of physics informed neural networks (PINNs) for approximating PDEs. *arXiv preprint arXiv:2006.16144*, 2020.

Nash, J. Parabolic equations. *Proceedings of the National Academy of Sciences*, 43(8):754–758, 1957.

Nash, J. Continuity of solutions of parabolic and elliptic equations. *American Journal of Mathematics*, 80(4):931–954, 1958.

Payne, L. E. and Weinberger, H. F. An optimal Poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960.

Poincaré, H. Sur les équations aux dérivées partielles de la physique mathématique. *American Journal of Mathematics*, pp. 211–294, 1890.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics informed deep learning (Part I): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

Schmidt, M. and Lipson, H. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

Sirignano, J. and Spiliopoulos, K. DGM: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375:1339–1364, 2018.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

Weller, H., Browne, P., Budd, C., and Cullen, M. Mesh adaptation on the sphere using optimal transport and the numerical solution of a Monge–Ampère type equation. *Journal of Computational Physics*, 308:102–123, 2016.

Yarotsky, D. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

A. Proofs from Section 6.1: Convergence Rate of Sequence

A.1. Proof of Lemma 4

Proof. In order to prove part 1, we will use the following integration by parts identity, for functions $r : \Omega \to \mathbb{R}$ such that and $s : \Omega \to \mathbb{R}$, and $r, s \in H_0^1(\Omega)$,

$$\int_\Omega \frac{\partial r}{\partial x_i} s dx = -\int_\Omega r \frac{\partial s}{\partial x_i} dx + \int_{\partial\Omega} rsnd\Gamma \tag{13}$$

where n_i is a normal at the boundary and $d\Gamma$ is an infinitesimal element of the boundary $\partial\Omega$.

Using the formula in (13) for functions $u, v \in H_0^1(\Omega)$, we have

$$\begin{aligned}
\langle D\mathcal{E}(u), v\rangle_{L^2(\Omega)} &= \langle -\nabla_x \cdot \partial_{\nabla u} L(x, u, \nabla u) + \partial_u L(x, u, \nabla u), v\rangle_{L^2(\Omega)} \\
&= -\int_\Omega \nabla_x \cdot \partial_{\nabla u} L(x, u, \nabla u) v + \partial_u L(x, u, \nabla u) v \; dx \\
&= -\int_\Omega \sum_{i=1}^d \frac{\partial \left(\partial_{\nabla u} L(x, u, \nabla u)\right)_i}{\partial x_i} v + \partial_u L(x, u, \nabla u) v \; dx \\
&= \int_\Omega \sum_{i=1}^d \left(\partial_{\nabla u} L(x, u, \nabla u)\right)_i \frac{\partial v}{\partial x_i} dx + \int_\Omega \sum_{i=1}^d \left(\partial_{\nabla u} L(x, u, \nabla u)\right)_i v n_i dx + \int_\Omega \partial_u L(x, u, \nabla u) v \; dx \\
&= \int \partial_{\nabla u} L(\nabla u) \cdot \nabla v + \partial_u L(x, u, \nabla u) v \; dx
\end{aligned}$$

where in the last equality we use the fact that the function $v \in H_0^1(\Omega)$, thus $v(x) = 0, \forall x \in \partial\Omega$.

To prove part 2. first note from Part 1. we know that $\langle D\mathcal{E}(u) - D\mathcal{E}(v), u - v\rangle_{L^2(\Omega)}$ takes the following form,

$$\begin{aligned}
&\langle D\mathcal{E}(u) - D\mathcal{E}(v), u - v\rangle_{L^2(\Omega)} \\
&= \langle \partial_{\nabla u} L(x, u, \nabla u) - \partial_{\nabla v} L(x, v, \nabla v), \nabla u - \nabla v\rangle_{L^2(\Omega)} + \langle \partial_u L(x, u, \nabla u) - \partial_v L(x, v, \nabla v), u - v\rangle_{L^2(\Omega)}
\end{aligned} \tag{14}$$

We know that for $x \in \Omega$, we have

$$\nabla^2_{(u,\nabla u)} L(x, u, \nabla u) \leq \mathrm{diag}([\Lambda, \Lambda \mathbf{1}_d])$$

Note that $\nabla_{(u,\nabla u)} L(x, u, \nabla u)$ is a vector, and we can write, $\partial_{(u,\nabla u)} L(x, u, \nabla u) = [\partial_u L(x, u, \nabla u), \partial_{\nabla u} L(x, u, \nabla u)]$ (here for two vectors a, b we define a new vector $c := [a, b]$ as their concatenation).

Using the smoothness of L can write,

$$\begin{aligned}
&[\partial_u L(x, u, \nabla u) - \partial_u L(x, v, \nabla v), \partial_{\nabla u} L(x, u, \nabla u) - \partial_{\nabla u} L(x, v, \nabla v)]^T \left([u - v, \nabla u - \nabla v]\right) \\
&\leq [u - v, \nabla u - \nabla v]^T \left(\mathrm{diag}([\Lambda, \Lambda \mathbf{1}_d])\right) [u - v, \nabla u - \nabla v] \\
&\leq \Lambda [u - v, \nabla u - \nabla v]^T [u - v, \nabla u - \nabla v]
\end{aligned}$$

This implies that for $x \in \Omega$ we have

$$\begin{aligned}
&\left(\partial_{\nabla u} L(x, u(x), \nabla u(x)) - \partial_{\nabla u} L(x, v(x), \nabla v(x))\right)^T (\nabla u(x) - \nabla v(x)) \\
&\qquad + \left(\partial_u L(x, u(x), \nabla u(x)) - \partial_u L(x, v(x), \nabla v(x))\right)^T (u(x) - v(x)) \\
&\leq \Lambda \|\nabla u(x) - \nabla v(x)\|_2^2 + \Lambda \|u(x) - v(x)\|_2^2
\end{aligned}$$

Integrating over Ω on both sides we get

$$\begin{aligned}
&\langle \partial_{\nabla u} L(x, u, \nabla u) - \partial_{\nabla v} L(x, v, \nabla v), \nabla u - \nabla v\rangle_{L^2(\Omega)} + \langle \partial_u L(x, u, \nabla u) - \partial_v L(x, v, \nabla v), u - v\rangle_{L^2(\Omega)} \\
&\leq \Lambda \|\nabla u - \nabla v\|^2_{L^2(\Omega)} + \Lambda \|u - v\|^2_{L^2(\Omega)} \\
&\leq \Lambda (1 + C_p^2) \cdot \|u - v\|^2_{H_0^1(\Omega)}.
\end{aligned}$$

the Poincare inequaltiy from Theorem 2 in the final equation. Hence plugging this result in Equation 14 we have,

$$\langle D\mathcal{E}(u) - D\mathcal{E}(v), u - v\rangle_{L^2(\Omega)} \leq (\Lambda + C_p^2\Lambda)\|u - v\|^2_{H_0^1(\Omega)}$$

This proves the right hand side of the inequality in part 2.

To prove the left and side we use similar to the upper bound, using the convexity of the $L(x, \cdot, \cdot)\colon \mathbb{R} \times \mathbb{R}^d$, we can lower bound the following term,

$$\begin{aligned}
&[\partial_u L(x, u, \nabla u) - \partial_u L(x, v, \nabla v), \partial_{\nabla u} L(x, u, \nabla u) - \partial_{\nabla u} L(x, v, \nabla v)]^T ([u - v, \nabla u - \nabla v]) \\
&\geq [u - v, \nabla u - \nabla v]^T (\mathrm{diag}([0, \lambda \mathbf{1}_d])) [u - v, \nabla u - \nabla v] \\
&\geq \lambda(\nabla u - \nabla v)^T(\nabla u - \nabla v)
\end{aligned}$$

Therefore, for all $x \in \Omega$ we have

$$\begin{aligned}
&(\partial_{\nabla u} L(x, u(x), \nabla u(x)) - \partial_{\nabla u} L(x, v(x), \nabla v(x))^T (\nabla u(x) - \nabla v(x)) \\
&+ (\partial_u L(x, u(x), \nabla u(x)) - \partial_u L(x, v(x), \nabla v(x))^T (u(x) - v(x)) \\
&\geq \lambda\|\nabla u(x) - \nabla v(x)\|_2^2
\end{aligned}$$

Integrating over Ω on both sides we get

$$\begin{aligned}
&\langle \partial_{\nabla u} L(x, u, \nabla u) - \partial_{\nabla v} L(x, v, \nabla v), \nabla u - \nabla v\rangle_{L^2(\Omega)} \\
&+ \langle \partial_u L(x, u, \nabla u) - \partial_v L(x, v, \nabla v), u - v\rangle_{L^2(\Omega)} \\
&\geq \lambda\|\nabla u - \nabla v\|^2_{L^2(\Omega)} \\
&= \lambda\|u - v\|^2_{H_0^1(\Omega)}.
\end{aligned}$$

Therefore we have,

$$\lambda\|u - v\|^2_{H_0^1(\Omega)} \leq \langle D\mathcal{E}(u) - D\mathcal{E}(v), u - v\rangle_{L^2(\Omega)} \leq (\Lambda + C_p^2\Lambda)\|u - v\|^2_{H_0^1(\Omega)}$$

as we wanted.

To show part 3, we will again use the fact that the function for a given $x \in \Omega$ the function $L(x, \cdot, \cdot)$ is strongly convex and smooth. Therefore using Taylor's Theorem $L(x, u + v, \nabla u + \nabla v)$ along $L(x, u, \nabla u)$ we can re-write the energy function as:

$$\begin{aligned}
&\mathcal{E}(u + v) \\
&= \int_\Omega L(x, u(x) + v(x), \nabla u(x) + \nabla v(x)) - f(x)(u(x) + v(x))dx \\
&= \int_\Omega L(x, u(x), \nabla u(x)) + \nabla_{(u,\nabla u)} L(x, u(x), \nabla u(x))^T [v(x), \nabla v(x)] \\
&\qquad + \frac{1}{2}[v(x), \nabla v(x)]^T \nabla^2_{(u,\nabla u)} L(\tilde{x}, u(\tilde{x}), \nabla \tilde{x})[u(x), \nabla u(x)] - \int f(x)(u(x) + v(x))dx \\
&= \int_\Omega L(x, u(x), \nabla u(x)) + [\partial_u L(u, u(x), \nabla u(x)), \partial_{\nabla u} L(x, u(x), \nabla u(x))]^T [v(x), \nabla v(x)] \\
&\qquad + \frac{1}{2}[v(x), \nabla v(x)]^T \nabla^2_{(u,\nabla u)} L(\tilde{x}, u(\tilde{x}), \nabla u(\tilde{x}))[v(x), \nabla v(x)] - \int f(x)(u(x) + v(x))dx
\end{aligned} \tag{15}$$

From Equation 2 of Definition 1 we know that for a given $x \in \Omega$ the function $L(x, \cdot, \cdot)$ is smooth and convex. In particular we know that,

$$\mathrm{diag}([0, \lambda I_d]) \leq \nabla^2_{(u,\nabla u)} \leq \mathrm{diag}[\Lambda, \Lambda I_d].$$

Using this to upper bound (15) we get,

$$\begin{aligned}
\mathcal{E}(u+v) &\leq \int_\Omega L(x,u(x),\nabla u(x)) + [\partial_u L(u,u(x),\nabla u(x)), \partial_{\nabla u} L(x,u(x),\nabla u(x))]^T [v(x),\nabla v(x)] \\
&\quad + \frac{\Lambda}{2}[v(x),\nabla v(x)]^T[v(x),\nabla v(x)] - \int f(x)(u(x)+v(x))dx \\
&= \int_\Omega L(x,u(x),\nabla u(x)) + \partial_u L(u,u(x),\nabla u(x))v(x) + \partial_{\nabla u} L(x,u(x),\nabla u(x))\nabla v(x) \\
&\quad + \frac{\Lambda}{2}\left(v(x)^2 + \|\nabla v(x)\|_2^2\right) - \int f(x)(u(x)+v(x))dx \\
&= \mathcal{E}(u) + \langle D\mathcal{E}(u) - f, v\rangle_{L^2(\Omega)} + \frac{\Lambda}{2}\left(\|v\|_{L^2(\Omega)} + \|v\|_{H_0^1(\Omega)}\right) \\
\implies \mathcal{E}(u+v) &\leq \mathcal{E}(u) + \langle D\mathcal{E}(u) - f, v\rangle_{L^2(\Omega)} + \frac{\Lambda(1+C_p^2)}{2}\|v\|_{H_0^1(\Omega)} \qquad (16)
\end{aligned}$$

We can similarly lower bound (15) by using the convexity of $\nabla^2_{(u,\nabla u)} L$ as

$$\begin{aligned}
\mathcal{E}(u+v) &\geq \int_\Omega L(x,u(x),\nabla u(x)) + [\partial_u L(u,u(x),\nabla u(x)), \partial_{\nabla u} L(x,u(x),\nabla u(x))]^T [v(x),\nabla v(x)] \\
&\quad + \frac{\Lambda}{2}\nabla v(x)^T \nabla v(x) - \int f(x)(u(x)+v(x))dx \\
&= \int_\Omega L(x,u(x),\nabla u(x)) + \partial_u L(u,u(x),\nabla u(x))v(x) + \partial_{\nabla u} L(x,u(x),\nabla u(x))\nabla v(x) \\
&\quad + \frac{\lambda}{2}\|\nabla v(x)\|_2^2 - \int f(x)(u(x)+v(x))dx \\
\implies \mathcal{E}(u+v) &\geq \mathcal{E}(u) + \langle D\mathcal{E}(u) - f, v\rangle_{L^2(\Omega)} + \frac{\lambda}{2}\|v\|_{H_0^1(\Omega)} \qquad (17)
\end{aligned}$$

Combining (16) and (17) we get,

$$\frac{\lambda}{2}\|\nabla v\|^2_{L^2(\Omega)} + \langle D\mathcal{E}(u) - f, v\rangle_{L^2(\Omega)} \leq \mathcal{E}(u+v) - \mathcal{E}(u) \leq \langle D\mathcal{E}(u) - f, v\rangle_{L^2(\Omega)} + \frac{(1+C_p)^2\Lambda}{2}\|\nabla v\|^2_{L^2(\Omega)}$$

Finally, part 4 follows by plugging in $u = u^\star$ and $v = u - u^\star$ in part 3 and using the fact that $D\mathcal{E}(u^\star) = f$. □

A.2. Proof of Lemma 5

Proof. Let $\{\lambda_i, \phi_i\}_{i=1}^\infty$ denote the (eigenvalue, eigenfunction) pairs of the operator $-\Delta$ where $0 < \lambda_1 \leq \lambda_2 \leq \cdots$, which are real and countable. (Evans (2010), Theorem 1, Section 6.5)

Using the definition of eigenvalues and eigenfunctions, we have

$$\begin{aligned}
\lambda_1 &= \inf_{v \in H_0^1(\Omega)} \frac{\langle -\Delta v, v\rangle_{L^2(\Omega)}}{\|v\|^2_{L^2(\Omega)}} \\
&= \inf_{v \in H_0^1(\Omega)} \frac{\langle \nabla v, \nabla v\rangle_{L^2(\Omega)}}{\|v\|^2_{L^2(\Omega)}} \\
&= \frac{1}{C_p}.
\end{aligned}$$

where in the last equality we use Theorem 2.

Let us write the functions v, w in the eigenbasis as $v = \sum_i \mu_i \phi_i$. Notice that an eigenfunction of $-\Delta$ is also an eigenfunction for $(I - \Delta)^{-1}$, with correspondinding eigenvalue $\frac{1}{1+\lambda_i}$.

Thus, to show part 1, we have,

$$\begin{aligned}\left\|(I-\Delta)^{-1}\nabla_x\cdot\nabla v\right\|_{L^2(\Omega)}^2 &= \left\|(I-\Delta)^{-1}\Delta v\right\|_{L^2(\Omega)}^2\\ &= \left\|\sum_{i=1}^{\infty}\frac{\lambda_i}{1+\lambda_i}\mu_i\phi_i\right\|_{L^2(\Omega)}^2\\ &\leq \left\|\sum_{i=1}^{\infty}\mu_i\phi_i\right\|_{L^2(\Omega)}^2\\ &= \sum_{i=1}^{\infty}\mu_i^2 = \|u\|_{L^2(\Omega)}^2\end{aligned}$$

where in the last equality we use the fact that ϕ_i are orthogonal.

Now, bounding $\langle(I-\Delta)^{-1}v,v\rangle_{L^2(\Omega)}$ for part 2. we use the fact that eigenvalues of the operator $(I-\Delta)^{-1}$ are of the form $\left\{\frac{1}{1+\lambda_i}\right\}_{i=1}^{\infty}$ we have,

$$\begin{aligned}\langle(I-\Delta)^{-1}v,v\rangle_{L^2(\Omega)} &= \left\langle\sum_{i=1}^{\infty}\frac{\mu_i}{1+\lambda_i}\phi_i,\sum_{i=1}^{\infty}\mu_i\phi_i\right\rangle_{L^2(\Omega)}\\ &\leq \left\langle\sum_{i=1}^{\infty}\mu_i\phi_i,\sum_{i=1}^{\infty}\mu_i\phi_i\right\rangle_{L^2(\Omega)}\\ &= \|u\|_{L^2(\Omega)}^2 \qquad (18)\end{aligned}$$

Before proving part 3., note that since $\lambda_1\leq\lambda_2\leq\cdots$ and $\frac{x}{1+x}$ is monotonically increasing, we have for all $i\in\mathbb{N}$

$$\frac{1}{1+\lambda_i}\geq\frac{1}{(1+C_p)\lambda_i} \qquad (19)$$

and note that $\frac{1}{\lambda_i}$ are the eigenvalues for $(-\Delta)^{-1}$ for all $i\in\mathbb{N}$. Using the inequality in (19) and the fact that ϕ_i's are orthogonal, we can further lower bound $\langle(I-\Delta)^{-1}v,v\rangle_{L^2(\Omega)}$ as follows,

$$\begin{aligned}\langle(I-\Delta)^{-1}v,v\rangle_{L^2(\Omega)} &= \sum_{i=1}^{\infty}\frac{\mu_i^2}{1+\lambda_i}\|\phi_i\|_{L^2(\Omega)}^2\\ &\geq \sum_{i=1}^{\infty}\frac{\mu_i^2}{(1+pc)\lambda_i}\|\phi_i\|_{L^2(\Omega)}^2\\ &= \frac{1}{1+C_p}\langle(-\Delta)^{-1}v,v\rangle_{L^2(\Omega)},\end{aligned}$$

where we use the following set of equalities in the last step,

$$\langle(-\Delta)^{-1}v,v\rangle_{L^2(\Omega)} = \left\langle\sum_{i=1}^{\infty}\frac{\mu_i}{\lambda_i}\phi_i,\sum_{i=1}^{\infty}\mu_i\phi_i\right\rangle_{L^2(\Omega)} = \sum_{i=1}^{\infty}\frac{\mu_i^2}{\lambda_i}\|\phi_i\|_{L^2(\Omega)}^2. \qquad \square$$

A.3. Proof of Lemma 6: Convergence of Preconditioned Gradient Descent

Proof. For the analysis we consider $\eta=\frac{\lambda^4}{4(1+C_p)^7\Lambda^4}$

Taylor expanding as in (16), we have

$$\mathcal{E}(u_{t+1}) \leq \mathcal{E}(u_t) - \eta \underbrace{\left\langle D\mathcal{E}(\nabla u_t) - f, (I-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - f\right)\right\rangle_{L^2(\Omega)}}_{\text{Term 1}} + \underbrace{\frac{\eta^2 (1+C_p)^2 \Lambda}{2} \left\|\nabla_x (I-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - f\right)\right\|^2_{L^2(\Omega)}}_{\text{Term 2}}. \tag{20}$$

where we have in (16) plugged in $u_{t+1} - u_t = -\eta\,(I-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - f\right)$.

First we lower bound *Term 1*. Since $u^\star$ is the solution to the PDE in (4), we have $D\mathcal{E}(u^\star) = f$. Therefore we have

$$\left\langle D\mathcal{E}(u_t) - f, (I-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - f\right)\right\rangle_{L^2(\Omega)} = \left\langle D\mathcal{E}(u_t) - D\mathcal{E}(u^\star), (I-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right\rangle_{L^2(\Omega)} \tag{21}$$

Using the result from Lemma 5 part 3., we have,

$$\begin{aligned}
&\langle D\mathcal{E}(u_t) - D\mathcal{E}(u^\star), (I-\Delta_x)^{-1} D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\rangle_{L^2(\Omega)} \\
&\geq \frac{1}{1+C_p}\left(\langle D\mathcal{E}(u_t) - D\mathcal{E}(u^\star), (-\Delta_x)^{-1} D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\rangle_{L^2(\Omega)}\right)
\end{aligned}$$

Using the Equation (21) and the fact that $\langle D\mathcal{E}(u), v\rangle_{L^2(\Omega)} = \langle \partial_{\nabla u} L(x,u,\nabla u), \nabla v\rangle_{L^2(\Omega)} + \langle \partial_u L(x,u,\nabla u), v\rangle_{L^2(\Omega)}$ from Lemma 4 we get,

$$\begin{aligned}
&\langle D\mathcal{E}(u_t) - D\mathcal{E}(u^\star), (I-\Delta_x)^{-1} D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\rangle_{L^2(\Omega)} \\
&\geq \frac{1}{1+C_p}\left(\langle D\mathcal{E}(u_t) - D\mathcal{E}(u^\star), (-\Delta_x)^{-1} D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\rangle_{L^2(\Omega)}\right) \\
&= \frac{1}{1+C_p}\left(\langle \partial_{\nabla u} L(x,u_t,\nabla u_t) - \partial_{\nabla u} L(x,u^\star,\nabla u^\star), \nabla_x (-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\rangle_{L^2(\Omega)}\right) \\
&\quad + \frac{1}{1+C_p}\left(\langle \partial_u L(x,u_t,\nabla u_t) - \partial_u L(x,u^\star,\nabla u^\star), (-\Delta_x)^{-1}(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star))\rangle_{L^2(\Omega)}\right) \\
&= \frac{1}{1+C_p}\Big\langle \nabla_{(u,\nabla u)} L(x,u_t,\nabla u_t) - \nabla_{(u,\nabla u)} L(x,u^\star,\nabla u^\star), \\
&\qquad \left[(-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right), \nabla_x (-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right]\Big\rangle_{L^2(\Omega)}
\end{aligned} \tag{22}$$

where we combine the terms $\nabla_x\,(-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)$ and $\nabla_x (-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)$ into a single vector in the last step.

Now, note that since for any $x \in \Omega$ the function $L(x,\cdot,\cdot)$ is strongly convex, we have

$$\nabla^2_{(u,\nabla u)} L(x,\nabla u,\nabla x) \geq \mathrm{diag}([0,\lambda \mathbf{1}_d])$$

Therefore for all x we can bound $\nabla_{(u,\nabla u)} L(x,u_t(x),\nabla u_t(x)) - \nabla_{(u,\nabla u)} L(x,u^\star(x),\nabla u^\star(x))$

$$\begin{aligned}
&\nabla_{(u,\nabla u)} L(x,u_t(x),\nabla u_t(x)) - \nabla_{(u,\nabla u)} L(x,u^\star(x),\nabla u^\star(x)) \\
&= [u_t(x) - u^\star(x), \nabla u_t(x) - \nabla u^\star(x)]^T \left(\nabla^2_{(u,\nabla u)} L(\tilde{x},u(\tilde{x}),\nabla u(\tilde{x}))\right)
\end{aligned} \tag{23}$$

where $\tilde{x} \in \Omega$ (and potentially different from x).

Using (23) in (22), we can lower bound the term as follows:

$$
\begin{aligned}
&\langle D\mathcal{E}(u_t) - D\mathcal{E}(u^\star), (I-\Delta_x)^{-1} D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\rangle_{L^2(\Omega)} \\
&\geq \frac{1}{1+C_p}\Bigg\langle [u_t - u^\star, \nabla u_t - \nabla u^\star]^T \left(\nabla^2_{(u,\nabla u)} L(\tilde{x}, u(\tilde{x}), \nabla u(\tilde{x})\right), \\
&\qquad \left[(-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right), \nabla_x (-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right]\Bigg\rangle_{L^2(\Omega)} \\
&\geq \frac{1}{1+C_p}\left\langle [0, \lambda\left(\nabla u_t(x) - \nabla u^\star(x)\right)], \left[(-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right), \nabla_x (-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right]\right\rangle_{L^2(\Omega)} \\
&= \frac{\lambda}{1+C_p}\left\langle \nabla u_t - \nabla u^\star, \nabla_x (-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right\rangle_{L^2(\Omega)} \\
&\overset{(i)}{=} \frac{\lambda}{1+C_p}\left\langle (-\Delta)u_t - (-\Delta)u^\star, (-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right\rangle_{L^2(\Omega)} \\
&\overset{(ii)}{=} \frac{\lambda}{1+C_p}\left\langle (-\Delta)^{-1}(-\Delta)u_t - (-\Delta)^{-1}(-\Delta)u^\star, \left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right\rangle_{L^2(\Omega)} \\
&\overset{(iii)}{=} \frac{\lambda}{1+C_p}\left\langle u_t - u^\star, \left(D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\right)\right\rangle_{L^2(\Omega)} \\
&\overset{(iv)}{\geq} \frac{\lambda^2}{1+C_p}\|u_t - u^\star\|^2_{H^1_0(\Omega)}
\end{aligned}
\tag{24}
$$

Here, we use the fact that for all $u, v \in H^1_0(\Omega)$ we have $\langle \nabla u, \nabla v\rangle_{L^2(\Omega)} = \langle -\Delta u, v\rangle_{L^2(\Omega)}$, i.e., Green's identity (along with the fact that we have a Dirichlet Boundary condition) to get step (i). We use the symmetry of the operator $(-\Delta)^{-1}$ in step (ii), and the fact that for a function $g \in H^1_0(\Omega)$ $(-\Delta)^{-1}(-\Delta)g = g$ in step (iii). We finally use Part 2 of Lemma 4 in the final step.

Hence finally Term 1 can be simplified as,

$$
\begin{aligned}
&\langle D\mathcal{E}(u_t) - D\mathcal{E}(u^\star), (I-\Delta_x)^{-1} D\mathcal{E}(u_t) - D\mathcal{E}(u^\star)\rangle_{L^2(\Omega)} \\
&\geq \frac{\lambda^2}{1+C_p}\|u_t - u^\star\|^2_{H^1_0(\Omega)} \\
&\geq \frac{2\lambda^2}{(1+C_p)^3\Lambda}\left(\mathcal{E}(u_t) - \mathcal{E}(u^\star)\right)
\end{aligned}
$$

where we use Part 4 from Lemma 4 in the final step.

We will proceed to upper bounding *Term 2*. Using the definition of $H^1_0(\Omega)$ norm, we can re-write *Term 2* as,

$$
\left\|\nabla_x (1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - f\right)\right\|^2_{L^2(\Omega)} = \left\|(1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t) - f\right)\right\|^2_{H^1_0(\Omega)}
$$

Writing the $H_0^1(\Omega)$ norm in its variational form (since $H_0^1(\Omega)$ norm is self-adjoint, Lemma 16) and upper bounding it,

$$\begin{aligned}
&\left\|(1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t)-f\right)\right\|_{H_0^1(\Omega)} \\
&= \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \left\langle \nabla_x (1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t)-f\right), \nabla v\right\rangle_{L^2(\Omega)} \\
&= \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \left\langle \nabla_x (1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t)-D\mathcal{E}(u^\star)\right), \nabla v\right\rangle_{L^2(\Omega)} \\
&\overset{(i)}{=} \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \left\langle (1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t)-D\mathcal{E}(u^\star)\right), -\Delta v\right\rangle_{L^2(\Omega)} \\
&\overset{(ii)}{=} \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \left\langle (-\Delta)(1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t)-D\mathcal{E}(u^\star)\right), v\right\rangle_{L^2(\Omega)} \\
&\leq \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \langle D\mathcal{E}(u_t)-D\mathcal{E}(u^\star), v\rangle_{L^2(\Omega)}
\end{aligned} \tag{25}$$

here, step (i) follows from the equality that for all $u, v \in H_0^1(\Omega)$ we have $\langle \nabla u, \nabla v\rangle_{L^2(\Omega)} = \langle -\Delta u, v\rangle_{L^2(\Omega)}$ and the fact that $-\Delta$ is a symmetric operator in step (ii).

Finally we use Lemma 5 Part 1 for the final step. More precisely, we use Part 1 of Lemma 5 as follows, where for a $g \in H_0^1(\Omega)$ we can write,

$$\sup_{\substack{v\in L^2(\Omega)\\ \|v\|_{L^2(\Omega)}=1}} \langle(-\Delta)(I-\Delta)^{-1}g, v\rangle_{L^2(\Omega)} = \|-\Delta(I-\Delta)^{-1}g\|_{L^2(\Omega)} \leq \|g\|_{L^2(\Omega)} =: \sup_{\substack{v\in L^2(\Omega)\\ \|v\|_{L^2(\Omega)}=1}} \langle g, v\rangle_{L^2(\Omega)}$$

Note that, from Lemma 4 we know that for all u, v we can write the inner product $\langle D\mathcal{E}(u), v\rangle$ as follows

$$\begin{aligned}
\langle D\mathcal{E}(u), v\rangle_{L^2(\Omega)} &= \langle \partial_{\nabla u} L(x,u,\nabla u), v\rangle_{L^2(\Omega)} + \langle \partial_u L(x,u,\nabla u), v\rangle_{L^2(\Omega)} \\
&= \langle \nabla_{(u,\nabla u)} L(x,u,\nabla u), [v, \nabla v]\rangle_{L^2(\Omega)}
\end{aligned}$$

that is, we we combine $\partial_{\nabla u} L$ and $\partial_u L$ into a single vector $\nabla_{(u,\nabla u)} L := [\partial_u L(x,u,\nabla u), \partial_{\nabla u} L(x,u,\nabla u)] \in \mathbb{R}^{d+1}$ and combining u and ∇u as a vector $[u, \nabla u]$.

Using this form and re-writing (25) and using the fact that for $x \in \Omega$ $L(x, \cdot, \cdot)$ is convex and smooth in step (i), we have

$$\begin{aligned}
&\left\|(1-\Delta_x)^{-1}\left(D\mathcal{E}(u_t)-f\right)\right\|_{H_0^1(\Omega)} \\
&\leq \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \left\langle \nabla_{(u,\nabla u)}L(x,u_t,\nabla u_t)-\nabla_{(u,\nabla u)}L(x,u^\star,\nabla u^\star),[v,\nabla v]\right\rangle_{L^2(\Omega)} \\
&\overset{(i)}{=} \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \left\langle [u_t-u^\star,\nabla u_t-\nabla u^\star]^T\nabla^2_{(u,\nabla u)}L(\tilde{x},u(\tilde{x}),\nabla u(\tilde{x})),[v,\nabla v]\right\rangle_{L^2(\Omega)} \\
&\leq \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \Lambda\left\langle [u_t-u^\star,\nabla u_t-\nabla u^\star]^T,[v,\nabla v]\right\rangle_{L^2(\Omega)} \\
&= \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \Lambda\left\langle u_t-u^\star,v\right\rangle_{L^2(\Omega)}+\Lambda\left\langle \nabla(u_t-u^\star),\nabla v\right\rangle_{L^2(\Omega)} \\
&= \sup_{\substack{v\in H_0^1(\Omega)\\ \|v\|_{H_0^1(\Omega)}=1}} \Lambda C_p^2\|u_t-u^\star\|_{H_0^1(\Omega)}\|v\|_{H_0^1(\Omega)}+\Lambda\|u_t-u^\star\|_{H_0^1(\Omega)}\|v\|_{H_0^1(\Omega)} \\
&= \Lambda(1+C_p^2)\|u_t-u^\star\|_{H_0^1(\Omega)} \leq \Lambda(1+C_p)^2\|u_t-u^\star\|_{H_0^1(\Omega)}
\end{aligned} \tag{26}$$

where we use the Poincare Inequality 2 in the final step.

Therefore, from the final result in (26) we can upper bound Term 2 in (20) to get,

$$\begin{aligned}
\left\|\nabla_x(I-\Delta_x)^{-1}D\mathcal{E}(u_t)\right\|^2_{L^2(\Omega)} &\leq \Lambda^2(1+C_p)^2\|u_t-u^\star\|^2_{H_0^1(\Omega)} \\
&\leq \frac{\Lambda^2(1+C_p)^2}{\lambda}\left(\mathcal{E}(u_t)-\mathcal{E}(u^\star)\right)
\end{aligned}$$

where we use the result from part 4 from Lemma 4.

$$\implies \mathcal{E}(u_{t+1})-\mathcal{E}(u^\star) \leq \mathcal{E}(u_t)-\mathcal{E}(u^\star)-\left(\frac{2\lambda^2}{(1+C_p)^3\Lambda}-\eta\frac{(1+C_p)^4\Lambda^3}{\lambda}\right)\eta\left(\mathcal{E}(u_t)-\mathcal{E}(u^\star)\right)$$

Since $\eta = \frac{\lambda^4}{4(1+C_p)^7\Lambda^4}$ we have

$$\begin{aligned}
&\mathcal{E}(u_{t+1})-\mathcal{E}(u^\star) \leq \mathcal{E}(u_t)-\mathcal{E}(u^\star)-\frac{\lambda^2}{(1+C_p)^3\Lambda}\eta\left(\mathcal{E}(u_t)-\mathcal{E}(u^\star)\right) \\
&\implies \mathcal{E}(u_{t+1})-\mathcal{E}(u^\star) \leq \left(1-\frac{\lambda^6}{(1+C_p)^{10}\Lambda^5}\right)^t\left(\mathcal{E}(u_0)-\mathcal{E}(u^\star)\right).
\end{aligned}$$

□

B. Error Analysis

B.1. Proof of Lemma 11

Proof. We define for all t $r_t = \tilde{u}_t - u_t$, and will iteratively bound $\|r_t\|_{L^2(\Omega)}$.

Starting with $u_0 = 0$ and $\tilde{u}_t = 0$, we define the iterative sequences as,

$$\begin{cases} u_0 = u_0 \\ u_{t+1} = u_t - \eta(I-\Delta_x)^{-1}D\mathcal{E}(u_t) \end{cases}$$

$$\begin{cases} \tilde{u}_t = u_0 \\ \tilde{u}_{t+1} = \tilde{u}_t - \eta(I - \Delta_x)^{-1} D\tilde{\mathcal{E}}(\tilde{u}_t) \end{cases}$$

where $\eta \in \left(0, \frac{\lambda^4}{4(1+C_p)^7\Lambda^4}\right]$. Subtracting the two we get,

$$\begin{aligned} \tilde{u}_{t+1} - u_{t+1} &= \tilde{u}_t - u_t - \eta(I - \Delta_x)^{-1}\left(D\tilde{\mathcal{E}}(\tilde{u}_t) - D\mathcal{E}(u_t)\right) \\ \implies r_{t+1} &= r_t - \eta(I - \Delta_x)^{-1}\left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t)\right) \end{aligned} \tag{27}$$

Taking $H_0^1(\Omega)$ norm on both sides we get,

$$\|r_{t+1}\|_{H_0^1(\Omega)} \leq \|r_t\|_{H_0^1(\Omega)} + \eta \left\|(I - \Delta_x)^{-1}\left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t)\right)\right\|_{H_0^1(\Omega)} \tag{28}$$

Towards bounding $\left\|(I - \Delta_x)^{-1} D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t)\right\|_{H_0^1(\Omega)}$, from Lemma 15 we know that the dual norm of $\|w\|_{H_0^1(\Omega)}$ is $\|w\|_{H_0^1(\Omega)}$, thus,

$$\begin{aligned} &\left\|(I - \Delta_x)^{-1} D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t)\right\|_{H_0^1(\Omega)} \\ &= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \nabla(I - \Delta_x)^{-1}\left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t)\right), \nabla\varphi \right\rangle_{L^2(\Omega)} \\ &= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \nabla(I - \Delta_x)^{-1}\left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t + r_t)\right), \nabla\varphi \right\rangle_{L^2(\Omega)} \\ &\quad + \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \nabla(I - \Delta_x)^{-1}\left(D\mathcal{E}(u_t + r_t) - D\mathcal{E}(u_t)\right), \nabla\varphi \right\rangle_{L^2(\Omega)} \\ &= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle (I - \Delta_x)^{-1}\left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t + r_t)\right), \Delta\varphi \right\rangle_{L^2(\Omega)} \\ &\quad + \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle (I - \Delta_x)^{-1}\left(D\mathcal{E}(u_t + r_t) - D\mathcal{E}(u_t)\right), \Delta\varphi \right\rangle_{L^2(\Omega)} \\ &= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t + r_t)\right), (I - \Delta)^{-1}\Delta\varphi \right\rangle_{L^2(\Omega)} \\ &\quad + \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \left(D\mathcal{E}(u_t + r_t) - D\mathcal{E}(u_t)\right), (I - \Delta)^{-1}\Delta\varphi \right\rangle_{L^2(\Omega)} \\ &\leq \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t + r_t)\right), \varphi \right\rangle_{L^2(\Omega)} \\ &\quad + \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \left(D\mathcal{E}(u_t + r_t) - D\mathcal{E}(u_t)\right), \varphi \right\rangle_{L^2(\Omega)} \end{aligned} \tag{29}$$

Now from Assumption 1, we know that for all $x \in \Omega$ and $u \in H_0^1(\Omega)$ we have the following bounds on the difference of partials of L and $\tilde{L}$:

$$\sup \left\|\partial_u \tilde{L}(x, u(x), \nabla u(x)) - \partial_u L(x, u(x), \nabla u(x))\right\|_2 \leq \epsilon_L \|u(x)\|_2, \tag{30}$$

and

$$\sup \left\| \partial_{\nabla u} \tilde{L}(x, u(x), \nabla u(x)) - \partial_{\nabla u} L(x, u(x), \nabla u(x)) \right\|_2 \leq \epsilon_L \|u(x)\|_2, \tag{31}$$

Therefore, note that we can bound the difference of $\nabla_{(u,\nabla u)}\tilde{L}$ and $\nabla_{(u,\nabla u)}L$ for all $x \in \Omega$ and $u \in H_0^1(\Omega)$ as follows,

$$\begin{aligned}
&\sup \left\| \nabla_{(u,\nabla u)} \tilde{L}(x, u(x), \nabla u(x)) - \nabla_{(u,\nabla u)} L(x, u(x), \nabla u(x)) \right\|_2 \\
&\leq \sup \left\| \partial_{\nabla u} \tilde{L}(x, u(x), \nabla u(x)) - \partial_{\nabla u} L(x, u(x), \nabla u(x)) \right\|_2 + \sup \left\| \partial_{\nabla u} \tilde{L}(x, u(x), \nabla u(x)) - \partial_{\nabla u} L(x, u(x), \nabla u(x)) \right\|_2 \\
&\leq 2\epsilon_L \|u(x)\|_2
\end{aligned} \tag{32}$$

Note that, from Lemma 4 we know that for all u, v we can write the inner product $\langle D\mathcal{E}(u), v\rangle$ as follows

$$\begin{aligned}
\langle D\mathcal{E}(u), v\rangle_{L^2(\Omega)} &= \langle \partial_{\nabla u} L(x, u, \nabla u), v\rangle_{L^2(\Omega)} + \langle \partial_u L(x, u, \nabla u), v\rangle_{L^2(\Omega)} \\
&= \langle \nabla_{(u,\nabla u)} L(x, u, \nabla u), [v, \nabla v]\rangle_{L^2(\Omega)}
\end{aligned} \tag{33}$$

that is, we we combine $\partial_{\nabla u} L$ and $\partial_u L$ into a single vector $\nabla_{(u,\nabla u)} L := [\partial_u L(x, u, \nabla u), \partial_{\nabla u} L(x, u, \nabla u)] \in \mathbb{R}^{d+1}$ and combining u and ∇u as a vector $[u, \nabla u]$.

Using upper bound in Equation 32 we can upper bound $\sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t + r_t) \right), \varphi \right\rangle_{L^2(\Omega)}$ (by expanding it as in Equation 33) as follows,

$$\begin{aligned}
&\sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \left(D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t + r_t) \right), \varphi \right\rangle_{L^2(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \nabla_{(u,\nabla u)} \tilde{L}(x, u_t + r_t, \nabla u_t + \nabla r_t) - \nabla_{(u,\nabla u)} L(x, u_t + r_t, \nabla u_t + \nabla r_t), [\varphi, \nabla \varphi] \right\rangle_{L^2(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \partial_{\nabla u} \tilde{L}(x, u_t + r_t, \nabla u_t + \nabla r_t) - \partial_{\nabla u} L(x, u_t + r_t, \nabla u_t + \nabla r_t), \nabla \varphi \right\rangle_{L^2(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \partial_{\nabla u} \tilde{L}(x, u_t + r_t, \nabla u_t + \nabla r_t) - \partial_{\nabla u} L(x, u_t + r_t, \nabla u_t + \nabla r_t), \varphi \right\rangle_{L^2(\Omega)} \\
&\quad + \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \partial_u \tilde{L}(x, u_t + r_t, \nabla u_t + \nabla r_t) - \partial_u L(x, u_t + r_t, \nabla u_t + \nabla r_t), \varphi \right\rangle_{L^2(\Omega)} \\
&\leq \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \epsilon_L \|u_t + r_t\|_{L^2(\Omega)} (1 + C_p) \|\varphi\|_{L^2(\Omega)} \\
&\leq \epsilon_L (1 + C_p) \|u_t + r_t\|_{L^2(\Omega)} \\
&\leq \epsilon_L (1 + C_p)^2 \|u_t + r_t\|_{H_0^1(\Omega)}
\end{aligned} \tag{34}$$

We can similarly bound $\sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \langle (D\mathcal{E}(u_t + r_t) - D\mathcal{E}(u_t)), \varphi\rangle_{L^2(\Omega)}$ where will use the convexity of the function $L(x, \cdot, \cdot)$ for all $u \in H_0^1(\Omega)$ to bound the gradient $\nabla_{(u,\nabla u)} L(x, u_t + r_t, \nabla u_t + \nabla r_t)$ using Taylor's theorem in the following way,

$$\begin{aligned}
&\nabla_{(u,\nabla u)} L(x, u_t + r_t, \nabla u_t + \nabla r_t) = \nabla_{(u,\nabla u)} L(x, u_t, \nabla u_t) + [r_t, \nabla r_t]^T \nabla^2_{(u,\nabla u)} L(\tilde{x}, u_t(\tilde{x}), \nabla u(\tilde{x})) \\
\implies &\nabla_{(u,\nabla u)} L(x, u_t + r_t, \nabla u_t + \nabla r_t) - \nabla_{(u,\nabla u)} L(x, u_t, \nabla u_t) = [r_t, \nabla r_t]^T \nabla^2_{(u,\nabla u)} L(\tilde{x}, u_t(\tilde{x}), \nabla u(\tilde{x}))
\end{aligned}$$

here $\tilde{x} \in \Omega$. Therefore, bounding $\sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \langle (D\mathcal{E}(u_t + r_t) - D\mathcal{E}(u_t)), \varphi \rangle_{L^2(\Omega)}$ we get,

$$\begin{aligned}
&\sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \langle (D\mathcal{E}(u_t + r_t) - D\mathcal{E}(u_t)), \varphi \rangle_{L^2(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle \nabla_{(u,\nabla u)} L(x, u_t + r_t, \nabla u_t + \nabla r_t) - \nabla_{(u,\nabla u)} L(x, u_t, \nabla u_t), [\varphi, \nabla \varphi] \right\rangle_{L^2(\Omega)} \\
&= \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \left\langle [r_t, \nabla r_t]^T \nabla^2_{(u,\nabla u)} L(\tilde{x}, u(\tilde{x}), \nabla u(\tilde{x})), [\varphi, \nabla \varphi] \right\rangle_{L^2(\Omega)} \\
&\leq \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \Lambda \left\langle [r_t, \nabla r_t]^T, [\varphi, \nabla \varphi] \right\rangle_{L^2(\Omega)} \\
&\leq \sup_{\substack{\varphi \in H_0^1(\Omega) \\ \|\varphi\|_{H_0^1(\Omega)}=1}} \Lambda \left(\|r_t\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)} + \|\nabla r_t\|_{L^2(\Omega)} \|\nabla \varphi\|_{L^2(\Omega)} \right) \\
&\leq \Lambda \left(1 + C_p\right)^2 \|r\|_{H_0^1(\Omega)}
\end{aligned} \tag{35}$$

Plugging in Equations (34) and (35) in (29) we get,

$$\begin{aligned}
\left\| (I - \Delta_x)^{-1} D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t) \right\|_{H_0^1(\Omega)} &\leq \epsilon_L (1 + C_p)^2 \|u_t + r_t\|_{H_0^1(\Omega)} + \Lambda (1 + C_p)^2 \|r\|_{H_0^1(\Omega)} \\
&= (1 + C_p)^2 (\epsilon_L + \Lambda) \|r_t\|_{H_0^1(\Omega)} + \epsilon (1 + C_p)^2 \|u_t\|
\end{aligned} \tag{36}$$

Furthermore, from Lemma 6 we have for all $t \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{E}(u_{t+1}) - \mathcal{E}(u^\star) &\leq \left(1 - \frac{\lambda^6}{(1 + C_p)^8 \Lambda^5}\right)^t \mathcal{E}(u_0) \\
&\leq \mathcal{E}(u_0)
\end{aligned}$$

and

$$\begin{aligned}
\|u_t - u^\star\|_{H_0^1(\Omega)} &\leq \frac{2}{\lambda} \left(\mathcal{E}(u_t) - \mathcal{E}(u_0)\right) \\
&\leq \frac{2}{\lambda} \mathcal{E}(u_0)
\end{aligned}$$

Hence we have that for all $t \in \mathbb{N}$,

$$\|u_t\|_{H_0^1(\Omega)} \leq \|u^\star\|_{H_0^1(\Omega)} + \frac{2}{\lambda} \mathcal{E}(u_0) =: R.$$

Putting this all together, we have

$$\left\| (I - \Delta_x)^{-1} D\tilde{\mathcal{E}}(u_t + r_t) - D\mathcal{E}(u_t) \right\|_{H_0^1(\Omega)} \leq (1 + C_p)^2 (\epsilon_L + \Lambda) \|r_t\|_{H_0^1(\Omega)} + \epsilon_L (1 + C_p)^2 R \tag{37}$$

Hence using the result from (37) in (28) and unfolding the recursion, we get,

$$\begin{aligned}
&\|r_{t+1}\|_{H_0^1(\Omega)} \leq \left(1 + \eta (1 + C_p)^2 (\epsilon_L + \Lambda)\right) \|r_t\|_{H_0^1(\Omega)} + (1 + C_p)^2 \epsilon_L \eta R \\
&\implies \|r_{t+1}\|_{H_0^1(\Omega)} \leq \frac{(1 + C_p)^2 \epsilon_L \eta R}{\eta (1 + C_p)^2 (\epsilon_L + \Lambda)} \left(\left(1 + \eta (1 + C_p)^2 \left(\epsilon_L + \Lambda\right)\right)^t - 1 \right) \\
&\implies \|r_{t+1}\|_{H_0^1(\Omega)} \leq \frac{\epsilon_L R}{\epsilon_L + \Lambda} \left(\left(1 + \eta (1 + C_p)^2 \left(\epsilon_L + \Lambda\right)\right)^t - 1 \right)
\end{aligned} \tag{38}$$

as we needed. □

C. Proofs for Section 6.2: Bounding the Barron Norm

C.1. Proof of Lemma 7: Barron Norm Increase after One Update

Proof. Note that the update equation looks like,

$$\begin{aligned}\tilde{u}_{t+1} &= \tilde{u}_t - \eta(I-\Delta_x)^{-1} D\mathcal{E}(u_t) \\ &= \tilde{u}_t - \eta(I-\Delta_x)^{-1}\left(-\nabla\cdot\partial_{\nabla u}L(x,\tilde{u}_t,\nabla\tilde{u}_t) + \partial_u L(x,\tilde{u}_t,\nabla\tilde{u}_t) - f\right) \\ &= \tilde{u}_t - \eta(I-\Delta_x)^{-1}\left(-\sum_{i=1}^{d}\partial_i\partial_{\nabla u}L(x,\tilde{u}_t,\nabla\tilde{u}_t) + \partial_u L(x,\tilde{u}_t,\nabla\tilde{u}_t) - f\right)\end{aligned} \tag{39}$$

From Lemma 8 we have

$$\|\nabla\tilde{u}_t\|_{\mathcal{B}(\Omega)} = \max_{i\in[d]}\|\partial_i\tilde{u}_t\|_{\mathcal{B}(\Omega)} \le 2\pi W_t\|\tilde{u}_t\|_{\mathcal{B}(\Omega)} \tag{40}$$

This also implies that

$$\max\{\|\tilde{u}_t\|_{\mathcal{B}(\Omega)}, \|\nabla\tilde{u}_t\|_{\mathcal{B}(\Omega)}\} \le 2\pi W_t\|\tilde{u}_t\|_{\mathcal{B}(\Omega)}.$$

Note that since $\tilde{u}_t \in \Gamma_{W_t}$ we have $\nabla\tilde{u}_t \in \Gamma_{2\pi W_t}$ and $L(x,\tilde{u}_t,\nabla\tilde{u}_t) \in \Gamma_{2\pi k_{\tilde{L}} W_t}$ (from Assumption 1).

Therefore, we can bound the Barron norm as,

$$\begin{aligned}&\left\|(I-\Delta_x)^{-1}\left(-\sum_{i=1}^{d}\partial_i\partial_{\nabla u}L(x,\tilde{u}_t,\nabla\tilde{u}_t) + \partial_u L(x,\tilde{u}_t,\nabla\tilde{u}_t) - f\right)\right\|_{\mathcal{B}(\Omega)} \\ &\overset{(i)}{\le} \left\|-\sum_{i=1}^{d}\partial_i\partial_{\nabla u}L(x,\tilde{u}_t,\nabla\tilde{u}_t)\right\|_{\mathcal{B}(\Omega)} + \|\partial_u L(x,\tilde{u}_t,\nabla\tilde{u}_t)\|_{\mathcal{B}(\Omega)} + \|f\|_{\mathcal{B}(\Omega)} \\ &\overset{(ii)}{\le} d\,\|\partial_i\partial_{\nabla u}L(x,\tilde{u}_t,\nabla\tilde{u}_t)\|_{\mathcal{B}(\Omega)} + \|\partial_u L(x,\tilde{u}_t,\nabla\tilde{u}_t)\|_{\mathcal{B}(\Omega)} + \|f\|_{\mathcal{B}(\Omega)} \\ &\le dB_{\tilde{L}}2\pi k_{\tilde{L}}(2\pi W_t)^{p_{\tilde{L}}}\|u\|_{\mathcal{B}(\Omega)}^{p_{\tilde{L}}} + B_{\tilde{L}}(2\pi W_t)^{p_{\tilde{L}}}\|u\|_{\mathcal{B}(\Omega)}^{p_{\tilde{L}}} + \|f\|_{\mathcal{B}(\Omega)} \\ &\le (2\pi k_{\tilde{L}}d+1)B_{\tilde{L}}(2\pi W_t)^{p_{\tilde{L}}}\|u\|_{\mathcal{B}(\Omega)}^{p_{\tilde{L}}} + \|f\|_{\mathcal{B}(\Omega)}\end{aligned}$$

where we use the fact that for a function h, we have $\|(I-\Delta_x)^{-1}h\|_{\mathcal{B}(\Omega)} \le \|h\|_{\mathcal{B}(\Omega)}$ from Lemma 8 in (i) and the bound from (40) in (ii).

Using the result of *Addition* from Lemma 8 we have

$$\begin{aligned}\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} &\le \|\tilde{u}_t\|_{\mathcal{B}(\Omega)} + \eta\left(2\pi k_{\tilde{L}}d+1)B_{\tilde{L}}(2\pi W_t)^{p_{\tilde{L}}}\|u\|_{\mathcal{B}(\Omega)}^{p_{\tilde{L}}} + \|f\|_{\mathcal{B}(\Omega)}\right) \\ &\le (1+\eta(2\pi k_{\tilde{L}}d+1)B_{\tilde{L}}(2\pi W_t)^{p_{\tilde{L}}})\,\|\tilde{u}\|_{\mathcal{B}(\Omega)}^{p_{\tilde{L}}} + \eta\|f\|_{\mathcal{B}(\Omega)}\end{aligned}$$

□

C.2. Proof of Lemma 9: Final Barron Norm Bound

Proof. From Lemma 7 we have

$$\begin{aligned}\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} &\le \|\tilde{u}_t\|_{\mathcal{B}(\Omega)} + \eta\left((2\pi k_{\tilde{L}}d+1)B(2\pi W_t)^p\|u\|_{\mathcal{B}(\Omega)}^p + \|f\|_{\mathcal{B}(\Omega)}\right) \\ &\le (1+\eta(2\pi k_{\tilde{L}}d+1)B(2\pi W_t)^p)\,\|u\|_{\mathcal{B}(\Omega)}^p + \eta\|f\|_{\mathcal{B}(\Omega)}\end{aligned}$$

Denoting the constant $A = (1 + \eta(2\pi k_{\tilde{L}} d + 1)B(2\pi W_t)^p)$ we have

$$\begin{aligned}
\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} &= A\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)} + \eta\|f\|_{\mathcal{B}(\Omega)} \\
\log\left(\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)}\right) &= \log\left(A\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)} + \eta\|f\|_{\mathcal{B}(\Omega)}\right) \\
&= \log\left(A\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)}\left(1 + \frac{\eta\|f\|_{\mathcal{B}(\Omega)}}{A\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)}}\right)\right) \\
&\leq \log\left(A\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)}\left(1 + \frac{\eta\|f\|_{\mathcal{B}(\Omega)}}{\max\{1, A\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)}\}}\right)\right) \\
&= \log\left(A\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)}\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right) \\
&= \log\left(\|\tilde{u}_t\|^p_{\mathcal{B}(\Omega)}\right) + \log\left(A\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right) \\
&= r\log(\|\tilde{u}_t\|_{\mathcal{B}(\Omega)}) + \log\left(A\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right) \qquad (41)
\end{aligned}$$

The above equation is a recursion of the form

$$x_{t+1} \leq r x_t + c$$

which implies

$$x_{t+1} \leq c\frac{p^t - 1}{p - 1} + p^t x_0.$$

Therefore the final bound in (41) is,

$$\begin{aligned}
&\log\left(\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)}\right) \leq r\log(\|\tilde{u}_t\|_{\mathcal{B}(\Omega)}) + \log\left(A\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right) \\
\implies &\log\left(\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)}\right) \leq \frac{r^n - 1}{r - 1}\log\left(A\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right) + p^t\log(\|\tilde{u}_0\|_{\mathcal{B}(\Omega)}) \\
\implies &\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} \leq \left(A\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right)^{\frac{p^t-1}{p-1}}\|\tilde{u}_0\|^{p^t}_{\mathcal{B}(\Omega)} \\
\implies &\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} \leq \left((1 + \eta(2\pi k_{\tilde{L}} d + 1)B_{\tilde{L}}(2\pi W_t)^p)\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right)^{\frac{p^t-1}{p-1}}\|\tilde{u}_0\|^{p^t}_{\mathcal{B}(\Omega)} \\
\overset{(i)}{\implies} &\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} \leq \left((1 + \eta(2\pi k_{\tilde{L}} d + 1)B_{\tilde{L}}(2\pi k^t_{\tilde{L}} W_0)^p)\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right)^{\frac{p^t-1}{p-1}}\|\tilde{u}_0\|^{p^t}_{\mathcal{B}(\Omega)} \\
\overset{(ii)}{\implies} &\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} \leq \left((1 + \eta(2\pi k_{\tilde{L}} d + 1)B_{\tilde{L}}(2\pi k_{\tilde{L}} W_0))\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right)^{pt + \frac{p^t-1}{p-1}}\|\tilde{u}_0\|^{p^t}_{\mathcal{B}(\Omega)} \\
\implies &\|\tilde{u}_{t+1}\|_{\mathcal{B}(\Omega)} \leq \left((1 + \eta 2\pi k_{\tilde{L}} W_0(2\pi k d + 1)B_{\tilde{L}})\left(1 + \eta\|f\|_{\mathcal{B}(\Omega)}\right)\right)^{pt + \frac{p^t-1}{p-1}}\left(\max\{1, \|\tilde{u}_0\|^{p^t}_{\mathcal{B}(\Omega)}\}\right)
\end{aligned}$$

where we use the fact that $W_t = k^T_{\tilde{L}} W_0$ since $\tilde{u}_t \in \Gamma_{k^T_{\tilde{L}} W_0}$ in step (i) and use the property that $(1 + x^p) \leq (1 + x)^p$ since $x > 0$ in step (ii). □

C.3. Proof of Lemma 10

Lemma 12 (Lemma 10 restated)**.** *Let*

$$f(x) = \sum_{\alpha, |\alpha| \leq P}\left(A_\alpha \prod_{i=1}^d x_i^{\alpha_i}\right)$$

where α is a multi-index and $x \in \mathbb{R}^d$ and $A_\alpha \in \mathbb{R}$ is a scalar. If $g : \mathbb{R}^d \to \mathbb{R}^d$ is a function such that $g \in \Gamma_W$, then we have $f \circ g \in \Gamma_{PW}$ and the Barron norm can be bounded as,

$$\|f \circ g\|_{\mathcal{B}(\Omega)} \leq d^{P/2}\left(\sum_{\alpha, |\alpha|=1}^P |A_\alpha|^2\right)^{1/2}\|g\|^P_{\mathcal{B}(\Omega)}$$

Proof. Recall from Definition 6 we know that for a vector valued function $g : \mathbb{R}^d \to \mathbb{R}^d$, we have

$$\|g\|_{\mathcal{B}(\Omega)} = \max_{i\in[d]} \|g_i\|_{\mathcal{B}(\Omega)}.$$

Then, using Lemma 8, we have

$$\begin{aligned}
\|f(g)\|_{\mathcal{B}(\Omega)} &= \left\| \sum_{\alpha,|\alpha|=0}^{P} A_\alpha \prod_{i=1}^{d} g_i^{\alpha_i} \right\|_{\mathcal{B}(\Omega)} \\
&\leq \sum_{\alpha,|\alpha|=0}^{P} \left\| A_\alpha \prod_{i=1}^{d} g_i^{\alpha_i} \right\|_{\mathcal{B}(\Omega)} \\
&\leq \sum_{\alpha,|\alpha|=0}^{P} |A_\alpha| \left\| \prod_{i=1}^{d} g_i^{\alpha_i} \right\|_{\mathcal{B}(\Omega)} \\
&\leq \sum_{\alpha,|\alpha|=0}^{P} |A_\alpha| \left\| \prod_{i=1}^{d} g_i^{\alpha_i} \right\|_{\mathcal{B}(\Omega)} \\
&\leq \sum_{\alpha,|\alpha|=0}^{P} |A_\alpha| \left(\prod_{i=1}^{d} \|g_i^{\alpha_i}\|_{\mathcal{B}(\Omega)} \right) \\
&\leq \sum_{\alpha,|\alpha|=0}^{P} |A_\alpha| \left(\prod_{i=1}^{d} \|g_i\|_{\mathcal{B}(\Omega)}^{\alpha_i} \right) \\
&= \sum_{\alpha,|\alpha|=0}^{P} |A_\alpha| \left(\prod_{i=1}^{d} \|g_i\|_{\mathcal{B}(\Omega)}^{\alpha_i} \right) \\
&\leq \left(\sum_{\alpha,|\alpha|=0}^{P} |A_\alpha|^2 \right)^{1/2} \left(\sum_{\alpha,|\alpha|=1}^{P} \left(\prod_{i=1}^{d} \|g_i\|_{\mathcal{B}(\Omega)}^{\alpha_i} \right)^2 \right)^{1/2}
\end{aligned} \tag{42}$$

where we have repeatedly used Lemma 8 and Cauchy-Schwartz in the last line. Using the fact that for a multivariate function $g : \mathbb{R}^d \to \mathbb{R}^d$ we have for all $i \in [d]$

$$\|g\|_{\mathcal{B}(\Omega)} \geq \|g_i\|_{\mathcal{B}(\Omega)}.$$

Therefore, from (42) we get,

$$\begin{aligned}
\|f(g)\|_{\mathcal{B}(\Omega)} &\leq \left(\sum_{\alpha,|\alpha|=0}^{P} |A_\alpha|^2 \right)^{1/2} \left(\sum_{\alpha,|\alpha|=1}^{P} \left(\|g\|_{\mathcal{B}(\Omega)}^{\sum_{i=1}^{d} \alpha_i} \right)^2 \right)^{1/2} \\
&\leq \left(\sum_{\alpha,|\alpha|=0}^{P} |A_\alpha|^2 \right)^{1/2} \left(\sum_{\alpha,|\alpha|=1}^{P} \left(\|g\|_{\mathcal{B}(\Omega)}^{\alpha} \right)^2 \right)^{1/2} \\
&\leq d^{P/2} \left(\sum_{\alpha,|\alpha|=0}^{P} |A_\alpha|^2 \right)^{1/2} \|g\|_{\mathcal{B}(\Omega)}^{P}
\end{aligned}$$

Since the maximum power of the polynomial can take is P from Corollary 1 we will have $f \circ g \in \Gamma_{PW}$. □

C.4. Proof of Lemma 8: Barron Norm Algebra

The proof of Lemma 8 is fairly similar to the proof of Lemma 3.3 in (Chen et al., 2021)—the change stemming from the difference of the Barron norm being considered

Proof. We first show the result for *Addition* and bound $\|h_1 + h_2\|_{\mathcal{B}(\Omega)}$,

$$\begin{aligned}
\|g_1 + g_2\|_{\mathcal{B}(\Omega)} &= \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) \, |\widehat{g_1 + g_2}(\omega)| \\
&= \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) \, |\hat{g}_1(\omega) + \hat{g}_2(\omega)| \\
&\leq \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) \, |\hat{g}_1(\omega)| + \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) \, |\hat{g}_2(\omega)| \\
\implies \|h_1 + h_2\|_{\mathcal{B}(\Omega)} &\leq \|h_1\|_{\mathcal{B}(\Omega)} + \|h_2\|_{\mathcal{B}(\Omega)}.
\end{aligned}$$

For *Multiplication*, first note that multiplication of functions is equal to convolution of the functions in the frequency domain, i.e., for functions $g_1 : \mathbb{R}^d \to d$ and $g_2 : \mathbb{R}^d \to d$, we have,

$$\widehat{g_1 \cdot g_2} = \hat{g}_1 * \hat{g}_2 \tag{43}$$

Now, to bound the Barron norm for the multiplication of two functions,

$$\begin{aligned}
\|g_1 \cdot g_2\|_{\mathcal{B}(\Omega)} &= \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) |\widehat{g_1 \cdot g_2}(\omega)| \\
&= \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) |\hat{g}_1 * \hat{g}_2(\omega)| \\
&= \sum_{\omega \in \mathbb{N}^d} \sum_{z \in \mathbb{N}^d} (1 + \|\omega\|_2) \, |\hat{g}_1(z) \hat{g}_2(\omega - z)| \\
&\leq \sum_{\omega \in \mathbb{N}^d} \sum_{z \in \mathbb{N}^d} (1 + \|\omega - z\|_2 + \|z\|_2 + \|z\|_2 \|\omega - z\|_2) \, |\hat{g}_1(z) \hat{g}_2(\omega - z)|
\end{aligned}$$

Where we use $\|\omega\|_2 \leq \|\omega - z\|_2 + \|z\|_2$ and the fact that

$$\sum_{\omega} \sum_{z} \|z\|_2 \|\omega - z\|_2 |\hat{g}_1(z) \hat{g}_2(\omega - z)| > 0.$$

Collecting the relevant terms together we get,

$$\begin{aligned}
\|g_1 \cdot g_2\|_{\mathcal{B}(\Omega)} &\leq \sum_{\omega \in \mathbb{N}^d} \sum_{z \in \mathbb{N}^d} (1 + \|\omega - z\|_2) \cdot (1 + \|z\|_2) \, |\hat{g}_1(z)| \, |\hat{g}_2(\omega - z)| \\
&= ((1 + \|\omega\|_2) \hat{g}_1(\omega)) * ((1 + \|\omega\|_2) \hat{g}_2(\omega))
\end{aligned}$$

Hence using Young's convolution identity from Lemma 13 we have

$$\begin{aligned}
\|g_1 \cdot g_2\|_{\mathcal{B}(\Omega)} &\leq \left(\sum_{\omega \in \mathbb{R}^d} (1 + \|w\|_2) \hat{g}_1(\omega) d\omega \right) \left(\sum_{\omega \in \mathbb{R}^d} (1 + \|w\|_2) \hat{g}_2(\omega) d\omega \right) \\
\implies \|g_1 \cdot g_2\|_{\mathcal{B}(\Omega)} &\leq \|h_1\|_{\mathcal{B}(\Omega)} \|h_2\|_{\mathcal{B}(\Omega)}.
\end{aligned}$$

In order to show the bound for *Derivative*, since $h \in \Gamma_W$, there exists a function $g : \mathbb{R}^d \to \mathbb{R}$ such that,

$$g(x) = \sum_{\|\omega\|_\infty \leq W} e^{2\pi i \omega^T x} \hat{g}(\omega) d\omega$$

Now taking derivative on both sides we get,

$$\partial_j g(x) = \sum_{\|\omega\|_\infty \leq W} i e^{i \omega^T x} 2\pi \omega_j \hat{g}(\omega) \tag{44}$$

This implies that we can upper bound $|\widehat{\partial_i g}(\omega)|$ as

$$\begin{aligned} \widehat{\partial_j g}(\omega) &= i2\pi\omega_j \hat{g}(\omega) \\ \implies |\widehat{\partial_j g}(\omega)| &\leq 2\pi W |\hat{g}(\omega)| \end{aligned} \tag{45}$$

Hence we can bound the Barron norm of $\partial_j h$ as follows:

$$\begin{aligned} \|\partial_j g\|_{\mathcal{B}(\Omega)} &= \sum_{\|\omega\|_\infty \leq W} (1 + \|\omega\|_\infty) \, |\widehat{\partial_j g}(\omega)| d\omega \\ &\leq \sum_{\|\omega\|_\infty \leq W} (1 + \|\omega\|_\infty) |2\pi W \hat{g}(\omega)| d\omega \\ &\leq 2\pi W \sum_{\|\omega\|_\infty \leq W} (1 + \|\omega\|_\infty) |\hat{g}(\omega)| d\omega \\ &\leq 2\pi W \|h\|_{\mathcal{B}(\Omega)} \end{aligned}$$

In order to show the preconditioning, note that for functions $g, f : \Omega^d \to \mathbb{R}$, if $f = (I - \Delta)^{-1} g$ then we have then we have $(I - \Delta) f = g$. Furthermore, by Lemma 14 we have

$$(1 + \|\omega\|_2^2) \hat{f}(\omega) = \hat{g}(\omega) \implies \hat{f}(\omega) = \frac{\hat{g}(\omega)}{1 + \|\omega\|_2^2}. \tag{46}$$

Bounding $\|(I - \Delta)^{-1} f\|_{\mathcal{B}(\Omega)}$,

$$\begin{aligned} \|(I - \Delta)^{-1} g\|_{\mathcal{B}(\Omega)} &= \sum_{\omega \in \mathbb{N}^d} \frac{1 + \|\omega\|_2}{(1 + \|\omega\|_2^2)} \hat{g}(\omega) d\omega \\ &\leq \sum_{\omega \in \mathbb{N}^d} (1 + \|\omega\|_2) \hat{g}(\omega) d\omega \\ \implies \|(I - \Delta)^{-1} g\|_{\mathcal{B}(\Omega)} &\leq \|g\|_{\mathcal{B}(\Omega)}. \end{aligned}$$

□

Corollary 1. *Let $g : \mathbb{R}^d \to \mathbb{R}$ then for any $k \in \mathbb{N}$ we have $\|g^k\|_{\mathcal{B}(\Omega)} \leq \|g\|^k_{\mathcal{B}(\Omega)}$. Furthermore, if the function $g \in \Gamma_W$ then the function $g^k \in \Gamma_{kW}$.*

Proof. The result from $\|g^k\|_{\mathcal{B}(\Omega)}$ follows from the multiplication result in Lemma 8 and we can show this by induction. For $n = 2$, we have from Lemma 8 we have,

$$\|g^2\|_{\mathcal{B}(\Omega)} \leq \|g\|^2_{\mathcal{B}(\Omega)} \tag{47}$$

Assuming that we have for all n till $k - 1$ we have

$$\|g^n\|_{\mathcal{B}(\Omega)} \leq \|g\|^n_{\mathcal{B}(\Omega)} \tag{48}$$

for $n = k$ we get,

$$\|g^k\|_{\mathcal{B}(\Omega)} = \|g g^{k-1}\|_{\mathcal{B}(\Omega)} \leq \|g\|_{\mathcal{B}(\Omega)} \|g^{k-1}\|_{\mathcal{B}(\Omega)} \leq \|g\|^k_{\mathcal{B}(\Omega)}. \tag{49}$$

To show that for any k the function $g^k \in \Gamma_{kW}$, we write g^k in the Fourier basis. We have:

$$\begin{aligned} g^k(x) &= \prod_{j=1}^{k} \left(\sum_{\|\omega_j\|_\infty \leq W} \hat{g}(\omega_j) e^{2i\pi\omega_j^T x} d\omega_j \right) \\ &= \sum_{\|\omega\|_\infty \leq kW} \left(\sum_{\sum_{l=1}^k \omega_l = \omega} \prod_{j=1}^{k} \hat{g}(\omega_j) d\omega_1 \ldots d\omega_k \right) e^{i2\pi\omega^T k} d\omega \end{aligned}$$

In particular, the coefficients with $\|\omega\|_\infty > kW$ vanish, as we needed. □

Lemma 13 (Young's convolution identity)**.** *For functions $g \in L^p(\mathbb{R}^d)$ and $h \in L^q(\mathbb{R}^d)$ and*

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + 1$$

where $1 \leq p, q, r \leq \infty$ we have

$$\|f * g\|_r \leq \|g\|_p \|h\|_q.$$

Here $*$ denotes the convolution operator.

Lemma 14. *For a differentiable function $f : [0,1]^d \to \mathbb{R}$, such that $f \in L^1(\mathbb{R}^d)$ we have*

$$\widehat{\nabla f}(\omega) = i2\pi\omega\hat{f}(\omega)$$

D. Existence Uniqueness and Definition of the Solution

D.1. Proof of Existence and Uniqueness of Minima

Proof. The proof follows a similar sketch of that provided in (Fernández-Real & Ros-Oton, 2020) Chapter 3, Theorem 3.3.

We first show that the minimizer $u^\star$ of the energy functional $\mathcal{E}(u)$ exists.

Note that from Definition 1 we have for a fixed $x \in \Omega$ the function $L(x, \cdot, \cdot)$ is convex and smooth it has a unique minimum, i.e., there exists a $(y_L, z_L) \in \mathbb{R} \times \mathbb{R}^d$ such that for all $(y, z) \in \mathbb{R} \times \mathbb{R}^d$ we have $L(x, y, z) \geq L(x, y_L, z_L)$ and that $\nabla L(x, y_L, z_L) = 0$. Furthermore, using (2) from Definition 1 this also implies the following,

$$\lambda\|z - z_L\|_2^2 \leq L(x, y, z) - L(x, y_L, z_L) \leq \Lambda\left(\|y - y_L\|_2^2 + \|z - z_L\|_2^2\right).$$

Note we can (w.l.o.g) assume that for a fixed $x \in \Omega$ we have, $L(x, 0, 0) = 0$, and $\nabla_{y,z} L(x, 0, 0) = 0$ (we can redefine L as $\widetilde{L}(x, y, z) = L(x, y + y_L, z + z_L) - L(x, y_L, z_L)$ if necessary), hence the above equation can be simplified to,

$$\lambda\|z\|_2^2 \leq L(x, y, z) \leq \Lambda\left(\|y\|_2^2 + \|z\|_2^2\right), \quad \forall p \in \Omega \times \mathbb{R} \times \mathbb{R}^d. \tag{50}$$

Now, we define,

$$\mathcal{E}_\circ = \inf\left\{\int_\Omega L(x, v, \nabla v) - fv\, dx : x \in \Omega, v \in H_0^1(\Omega)\right\}$$

. Let us first show that $\mathcal{E}_\circ$ is finite. Indeed, using (50) for any $v \in H_0^1(\Omega)$ and $x \in \Omega$, we have

$$\begin{aligned}
\mathcal{E}(v) &= \int_\Omega L(x, v, \nabla v) - fv\, dx \\
&\leq \int_\Omega \Lambda\left(\|v(x)\|_2^2 + \|\nabla v(x)\|_2^2\right) + \|f(x)v(x)\|_2 dx \\
&\leq \Lambda\left(\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2\right) + \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}
\end{aligned}$$

and is thus finite.

Moreover, using (50) for all $v \in H_0^1(\Omega)$ and $x \in \Omega$, $\mathcal{E}(v)$ can be lower bounded as

$$\begin{aligned}
\mathcal{E}(v) &= \int_\Omega L(x, v, \nabla v) - fv\, dx \\
&\geq \int_\Omega \lambda\|\nabla v(x)\|_2 - \|f(x)v(x)\|_2\, dx \\
&\geq \lambda\|\nabla v\|_{L^2(\Omega)}^2 - \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\
&\geq \frac{\lambda}{2}\|\nabla v\|_{L^2(\Omega)}^2 + \left(\frac{\lambda}{2C_p} - \frac{1}{C}\right)\|v\|_{L^2(\Omega)}^2 - C\|f\|_{L^2(\Omega)}^2
\end{aligned} \tag{51}$$

for some large constant C so that $\lambda/2C_p - 1/C > 0$., where we have used the Poincare inequality (Theorem 2) and Cauchy-Schwarz inequality to get the last inequality.

Let $\{u_k\}$ where $u_k \in H_0^1(\Omega)\ \forall k$ define a minimizing sequence of function, that is, we have $\mathcal{E}(u_k) \to \mathcal{E}_\circ = \inf_v \mathcal{E}(v)$ as $k \to 0$. From 51 we have for all k

$$\frac{\lambda}{2}\|\nabla u_k\|^2_{L^2(\Omega)} + \left(\frac{\lambda}{2C_p} - \frac{1}{C}\right)\|u_k\|^2_{L^2(\Omega)} - C\|f\|^2_{L^2(\Omega)} \leq \mathcal{E}(u_k).$$

Therefore since $\mathcal{E}(u_k)$ is bounded, we have that $\|u_k\|_{H_0^1(\Omega)}$ is uniformly bounded, and thus we can extract a weakly convergent subsequence. With some abuse of notations, let us without loss of generality assume that $u_k \rightharpoonup u$.

We will now show that if $u_k \rightharpoonup u$,

$$\mathcal{E}(u) \leq \liminf_{k\to\infty} \mathcal{E}(u_k) = \mathcal{E}_\circ$$

and therefore conclude that the limit u is a minimizer. This property is also referred to as weak-lower semi-continuity of $\mathcal{E}$.

In order to show the weak-lower semicontinuity of $\mathcal{E}$ we define the following set,

$$\mathcal{A}(t) := \{v \in H_0^1(\Omega) : \mathcal{E}(v) \leq t\}.$$

Furthermore, note that the functional $\mathcal{E}(v)$ is convex in v (since the function L is convex and the term $f(x)v(x)$ is linear), and this also implies that the set $\mathcal{A}(t)$ is convex.

Further, for any sequence of functions $\{w_k\}$ where $w_k \in \mathcal{A}(t)$ such that $w_k \to w$ from Fatou's Lemma,

$$\mathcal{E}(w) = \int_\Omega L(x, w(x), \nabla w(x)) - f(x)w(x)dx \leq \liminf_{k\to\infty} \int_\Omega L(x, w_k(x), \nabla w_k(x)) - f(x)w_k(x)dx \leq t$$

hence we also have that the function $w \in \mathcal{A}(t)$. Therefore the set $\mathcal{A}(t)$ is closed (w.r.t $H_0^1(\Omega)$ norm), and it is convex. Since the set $A(t)$ is closed and convex (it is also weakly closed) therefore if $w_k \to w$ it also implies that $w_k \rightharpoonup w$ in $H_0^1(\Omega)$.

Hence, consider a weakly converging sequence in $H_0^1(\Omega)$, i.e., $w_k \rightharpoonup w$ and define

$$t^* := \liminf_{k\to\infty} \mathcal{E}(w_k)$$

Now, for any $\varepsilon > 0$, there exists a subsequence $w_{k_{j,\varepsilon}} \rightharpoonup w$ in $H_0^1(\Omega)$ and $\mathcal{E}_{w_{k_{j,\varepsilon}}} \leq t^\star + \varepsilon$, that is, $w_{k_{j,\varepsilon}} \in \mathcal{A}(t^* + \varepsilon)$. This this is true for all $\epsilon > 0$ this implies that $\mathcal{E}(w) \leq t^* = \liminf_{k\to 0} \mathcal{E}$. Hence the function $\mathcal{E}$ is lower-semi-continuous, and hence the minimizer exists!

Now to show that the minimum is unique. Note the function $\mathcal{E}$ is convex in u. We will prove that the minima is unique by contradiction.

Let $u, v \in H_0^1(\Omega)$ be two (distinct) minima of $\mathcal{E}$, i.e., we have, $\mathcal{E}(u) = \mathcal{E}_\circ$ and $\mathcal{E}(v) = \mathcal{E}_\circ$.

Now using the fact that the function $L : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is convex, and the minimality of $\mathcal{E}_\circ$, we have for all $x \in \Omega$ we have

$$\begin{aligned}
\mathcal{E}_\circ \leq \mathcal{E}\left(\frac{u+v}{2}\right) &= \int_\Omega L\left(x, \frac{u(x)+v(x)}{2}, \frac{\nabla u(x) + \nabla v(x)}{2}\right) + f(x)\frac{u(x)+v(x)}{2} \\
&= \int_\Omega L\left(\frac{x+x}{2}, \frac{u(x)+v(x)}{2}, \frac{\nabla u(x) + \nabla v(x)}{2}\right) + f(x)\frac{u(x)+v(x)}{2} \\
&\leq \int_\Omega \frac{1}{2}\left(L\left(x, u(x), \nabla u(x)\right) + u(x)\right) + \int_\Omega \frac{1}{2}\left(L\left(x, v(x), \nabla v(x)\right) + v(x)\right) \\
&\leq \frac{1}{2}\mathcal{E}(u) + \frac{1}{2}\mathcal{E}(v) \\
\implies \mathcal{E}_\circ \leq \mathcal{E}\left(\frac{u+v}{2}\right) &\leq \frac{1}{2}\mathcal{E}(u) + \frac{1}{2}\mathcal{E}(v) = \mathcal{E}_\circ.
\end{aligned}$$

The last inequality is a contradiction and therefore the minima is unique. □

D.2. Proof of Lemma 2: Nonlinear Elliptic Variational PDEs

Proof of Lemma 2. If the function $u^\star$ minimizes the energy functional in Definition 1 then we have for all $\epsilon \in \mathbb{R}$

$$\mathcal{E}(u) \leq \mathcal{E}(u + \epsilon\varphi)$$

where $\varphi \in C_c^\infty(\Omega)$. That is, we have a minima at $\epsilon = 0$ and taking a derivative w.r.t ϵ and using Taylor expansion we get,

$$\begin{aligned}
d\mathcal{E}[u](\varphi) &= \lim_{\epsilon\to 0} \frac{\mathcal{E}(u+\epsilon\varphi) - \mathcal{E}(u)}{\epsilon} = 0 \\
&= \lim_{\epsilon\to 0} \frac{\int_\Omega L(x, u+\epsilon\varphi, \nabla u + \epsilon\nabla\varphi) - f(x)\left(u(x) + \epsilon\varphi(x)\right) - L(x,u,\nabla u) + f(x)u(x)\, dx}{\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{\int_\Omega L(x, u+\epsilon\varphi, \nabla u) + \partial_{\nabla u} L(x, u+\epsilon\varphi, \nabla u) + r_1(x) - \epsilon f(x)\epsilon\varphi(x) - L(x,u,\nabla u)\, dx}{\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{\int_\Omega L(x,u,\nabla u) + \epsilon\partial_u L(x,u,\nabla u)\varphi + r_2(x)}{\epsilon} \\
&\quad + \lim_{\epsilon\to 0} \frac{\epsilon\partial_{\nabla u} L(x,u,\nabla u)\nabla\varphi + \epsilon^2\partial_u\partial_{\nabla u} L(x,u,\nabla u)\nabla\varphi\cdot\varphi + r_1(x) - \epsilon f(x)\epsilon\varphi(x) - L(x,u,\nabla u)\, dx}{\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{\int_\Omega \epsilon\partial_{\nabla u} L(x,u,\nabla u)\nabla\varphi + \epsilon\partial_u L(x,u,\nabla u)u + r_1(x) + r_2(x) - \epsilon f(x)\varphi(x)\, dx}{\epsilon}
\end{aligned} \tag{52}$$

where for all $x \in \Omega$ we have,

$$\begin{aligned}
|r_1(x)| &\leq \frac{\epsilon^2}{2} \sup_{y\in\Omega} \left| \left(\left(\nabla u(x)\right)^T \partial^2_{\nabla u} L(y, u+\epsilon\varphi, \nabla u)\nabla u(x) \right) \right| \\
&\leq \frac{\Lambda\epsilon^2}{2} \|\nabla u(x)\|_2^2
\end{aligned} \tag{53}$$

Similarity we have,

$$\begin{aligned}
|r_2(x)| &\leq \frac{\epsilon^2}{2} \sup_{y\in\Omega} \left| \partial_u L(y,u,\nabla u) u(x)^2 \right| \\
&\leq \frac{\Lambda\epsilon^2}{2} u(x)^2
\end{aligned} \tag{54}$$

Using results from (52) and (54) in Equation (53) and taking $\epsilon \to 0$, the derivative in the direction of φ is,

$$d\mathcal{E}[u](\varphi) = \lim_{\epsilon\to 0} \frac{\int_\Omega \partial_{\nabla u} L(x,u,\nabla u)\nabla\varphi + \partial_u L(x,u,\nabla u)u - f(x)\varphi(x)\, dx}{\epsilon}$$

Since $\epsilon \to 0$ the final derivative is of the form,

$$d\mathcal{E}[u](\varphi) = \int_\Omega \left(\partial_{\nabla u} L(x,u,\nabla u)\nabla\varphi + \partial_u L(x,u,\nabla v)\varphi - f\varphi \right) dx = 0. \tag{55}$$

We will now use the following integration by parts identity, for functions $r : \Omega \to \mathbb{R}$ such that and $s : \Omega \to \mathbb{R}$, and $r, s \in H_0^1(\Omega)$,

$$\int_\Omega \frac{\partial r}{\partial x_i} s dx = -\int_\Omega r \frac{\partial s}{\partial x_i} dx + \int_{\partial\Omega} rsnd\Gamma \tag{56}$$

where n_i is a normal at the boundary and $d\Gamma$ is an infinitesimal element of the boundary $\partial\Omega$.

Using the identity in (56) in (55) we get,

$$\begin{aligned}
d\mathcal{E}[u](\varphi) &= \int_\Omega \Big(\partial_{\nabla u} L(x,u,\nabla u)\nabla\varphi + +\partial_u L(x,u,\nabla v)\varphi - f\varphi \Big) dx \\
&= \int_\Omega \Big(\sum_{i=1}^d \left(\partial_{\nabla u} L(x,u,\nabla u)\right)_i \partial_i \varphi + \partial_u L(x,u,\nabla v)\varphi - f\varphi \Big) dx \\
&= \int_\Omega \Big(\sum_{i=1}^d -\partial_i \left(\partial_{\nabla u} L(x,u,\nabla u)\right)_i \varphi + \partial_u L(x,u,\nabla v)\varphi - f\varphi \Big) dx \\
&= \int_\Omega \Big(-\nabla_x \cdot \left(\partial_{\nabla u} L(x,u,\nabla u)\right)\varphi + \partial_u L(x,u,\nabla u)\varphi - f\varphi \Big) dx = 0 \\
\implies d\mathcal{E}[u](\varphi) &= \int_\Omega \Big(-\operatorname{div}_{\mathrm{x}} \left(\partial_{\nabla u} L(x,u,\nabla u)\right)\varphi + \partial_u L(x,u,\nabla u)\varphi - f\varphi \Big) dx = 0
\end{aligned}$$

That is the minima for the energy functional is reached at a u which solves the following PDE,

$$d\mathcal{E}(u) := -\operatorname{div}_{\mathrm{x}} \left(\partial_{\nabla u} L(x,u,\nabla u)\right) + \partial_u L(x,u,\nabla u) = f.$$

where we define $d\mathcal{E}(\cdot)$ as the operator $-\operatorname{div}_{\mathrm{x}} \left(\partial_{\nabla u} L(x,\cdot,\nabla\cdot)\right) + \partial_u L(x,\cdot,\nabla\cdot)$.

□

D.3. Proof of Lemma 3: Poincare constant of Unit Hypercube

Proof of Lemma 3. We use the fact that the Poincare constant is the smallest eigenvalue of Δ, i.e.,

$$\frac{1}{C_p} := \inf_{u \in L^2(\Omega)} \frac{\|\Delta u\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}}.$$

Note that the eigenfunctions of Δ for the domain $\Omega := [0,1]^d$ are defined as

$$\phi_\omega(x) = \prod_{i=1}^d \sin(\pi i \omega_i x_i), \quad \forall \omega \in \mathbb{N}^d \ \ \& \ \ x \in \Omega.$$

Furthermore, this also implies that for all $\omega \in \mathbb{N}^d$ we have,

$$\Delta\phi_\omega = \pi^2 \|\omega\|_2^2 \phi_\omega.$$

We can expand any function $u \in H_0^1(\Omega)$ in terms of ϕ_ω as $u(x) = \sum_{\omega\in\mathbb{N}^d} d_\omega \phi_\omega(x)$ where $d_\omega = \langle u, \phi_\omega \rangle_{L^2(\Omega)}$. Note that for all $x \in \Omega$, we have,

$$\Delta u(x) = \sum_{\omega\in\mathbb{N}^d} \pi^2 \|\omega\|_2^2 d_\omega \phi_\omega(x).$$

Taking square $L^2(\Omega)$ norm on both sides, we get,

$$\begin{aligned}
\|\Delta u\|_{L^2(\Omega)}^2 &= \pi^4 \left\| \sum_{\omega \in \mathbb{N}^d} \|\omega\|_2^2 d_\omega \phi_\omega \right\|_{L^2(\Omega)}^2 \\
&\overset{(i)}{\geq} \pi^4 d^2 \left\| \sum_{\omega \in \mathbb{N}^d} d_\omega \phi_\omega \right\|_{L^2(\Omega)}^2 \\
&\overset{(ii)}{=} \pi^4 d^2 \|u\|_{L^2(\Omega)}^2 \\
\implies \frac{\|\Delta u\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}} &\geq \pi^2 d
\end{aligned}$$

where we use the fact that $\|\omega\|_2 \geq \sqrt{d}$ (since $\forall i \in [d]$ we have $\omega_i \in \mathbb{N}$) in step (i), and use the orthogonality of $\{\phi_\omega\}_{\omega \in \mathbb{N}^d}$ in (ii). Moreover, it's easy to see that equality can be achieved by taking $u = \phi_{(1,1,\ldots,1)}$.

Hence the Poincare constant can be calculated as,

$$\begin{aligned}
\frac{1}{C_p} := \inf_{u \in L^2(\Omega)} \frac{\|\Delta u\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}} &= \pi^2 d \\
\implies C_p &= \frac{1}{\pi^2 d}.
\end{aligned}$$

□

E. Important Helper Lemmas

Lemma 15. *The dual norm of* $\|\cdot\|_{H_0^1(\Omega)}$ *is* $\|\cdot\|_{H_0^1(\Omega)}$.

Proof. If $\|u\|_*$ denotes the dual norm of $\|u\|_{H_0^1(\Omega)}$, by definition we have,

$$\begin{aligned}
\|u\|_* &= \sup_{\substack{v \in H_0^1(\Omega) \\ \|v\|_{H_0^1(\Omega)}=1}} \langle u, v \rangle_{H_0^1(\Omega)} \\
&= \sup_{\substack{v \in H_0^1(\Omega) \\ \|v\|_{H_0^1(\Omega)}=1}} \langle \nabla u, \nabla v \rangle_{L^2(\Omega)} \\
&\leq \sup_{\substack{v \in H_0^1(\Omega) \\ \|v\|_{H_0^1(\Omega)}=1}} \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\
&= \|\nabla u\|_{L^2(\Omega)}
\end{aligned}$$

where the inequality follows by Cauchy- Schwarz. On the other hand, equality can be achieved by taking $v = \frac{u}{\|\nabla u\|_2}$. Thus, $\|u\|_* = \|\nabla u\|_{L^2(\Omega)} = \|u\|_{H_0^1(\Omega)}$ as we wanted. □

E.1. Useful properties of Laplacian and Laplacian Inverse

Lemma 16. *The operator* $(-\Delta)^{-1}$ *is self-adjoint.*

Proof. Note that since the operator $(-\Delta)^{-1}$ is bounded, to show that it is self-adjoint, we only need to show that the operator is also symmetric, i.e., for all $u, v \in H_0^1(\Omega)$ we have

$$\langle (-\Delta)^{-1} u, v \rangle_{L^2(\Omega)} = \langle u, (-\Delta)^{-1} v \rangle_{L^2(\Omega)}.$$

To show this, we first show that the operator Δ is symmetric. i.e, we have

$$\langle -\Delta u, v\rangle_{L^2(\Omega)} = \langle u, -\Delta v\rangle_{L^2(\Omega)} \tag{57}$$

This is a direct consequence of the Green's Identity where for functions $u, v \in C_0^\infty$ the following holds,

$$\begin{aligned}\int_\Omega -(\Delta u)v dx &= \int_\Omega \nabla u \cdot \nabla v dx + \int_{\partial\Omega} \frac{\partial u}{\partial n} v d\Gamma \\ &= \int_\Omega \nabla u \cdot \nabla v dx \\ &= -\int_\Omega u \Delta v dx + \int_{\partial\Omega} \frac{\partial v}{\partial n} u d\Gamma\end{aligned}$$

where we use the fact that since $u, v \in H_0^1(\Omega)$ we have $u(x) = 0$ and $v(x) = 0$ for all $x \in \partial\Omega$.

Now, taking $\tilde{u} = -\Delta u$ and $\tilde{v} = (-\Delta)^{-1} v$ from Equation (57) we get,

$$\begin{aligned}\langle -\Delta u, v\rangle_{L^2(\Omega)} &= \langle u, \Delta v\rangle_{L^2(\Omega)} \\ \langle \tilde{u}, (-\Delta)^{-1}\tilde{v}\rangle_{L^2(\Omega)} &= \langle (-\Delta)^{-1}\tilde{u}, \tilde{v}\rangle_{L^2(\Omega)}.\end{aligned}$$

Hence we have that the operator $(-\Delta)^{-1}$ is symmetric and bounded and therefore is self-adjoint. □

Lemma 17. *Given a vector valued function $f : \mathbb{R}^d \to \mathbb{R}^d$, such that $f \in C^2$ the following identity holds,*

$$\nabla \mathrm{div_x}(f) = \mathrm{div_x}(\nabla f). \tag{58}$$

Proof. We first simplify the right hand side of Equation (58). Note that since $\nabla f : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a is a matrix valued function the divergence of ∇f is going to be vector valued. More precisely for all $x \in \Omega$, $-\mathrm{div_x}(\nabla f)$ is defined as

$$\begin{aligned}\mathrm{div_x}(\nabla f(x)) &= \left[\sum_{j=1}^d \partial_j [\nabla f(x)]_i\right]_{i=1}^d \\ &= \left[\sum_{j=1}^d \partial_j \partial_i f(x)\right]_{i=1}^d\end{aligned} \tag{59}$$

where for a vector valued function the notation $[g(x)]_i$ denotes its ith coordinate, and the notation $[g(x)]_{i=1}^d := (g(x)_1, g(x)_2, \cdots, g(x)_d)$ denotes a d dimensional vector.

Now, simplifying the left hand side, for all $x \in \Omega$ we get,

$$\begin{aligned}\nabla \mathrm{div_x}(f(x)) &= \nabla\left(\sum_{j=1}^d \partial_j f(x)\right) \\ &= \left[\partial_i\left(\sum_{j=1}^d \partial_j f(x)\right)\right]_{i=1}^d \\ &= \left[\left(\sum_{j=1}^d \partial_i \partial_j f(x)\right)\right]_{i=1}^d\end{aligned} \tag{60}$$

Since the term in (59) is equal to (60) we have $\nabla \mathrm{div_x}(f) = \mathrm{div_x}(\nabla f)$. □

Lemma 18. *For a function $g : \mathbb{R}^d \to \mathbb{R}$ such that $g \in C^3$ the following identity holds,*

$$\Delta \nabla g = \nabla \Delta g$$

Proof. The term $\Delta\nabla g$ can be simplified as follows,

$$\begin{aligned}\Delta\nabla g &= \Delta\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_d}\right)\\ &= \Delta\left[\frac{\partial f}{\partial x_i}\right]_{i=1}^{d}\\ &= \left[\Delta\frac{\partial f}{\partial x_i}\right]_{i=1}^{d}\\ &= \left[\sum_{j=1}^{d}\frac{\partial}{\partial x_j^2}\frac{\partial f}{\partial x_i}\right]_{i=1}^{d}\\ &= \left[\sum_{j=1}^{d}\frac{\partial^2 f}{\partial x_j^2\partial x_i}\right]_{i=1}^{d}\end{aligned} \tag{61}$$

Further, $\nabla\Delta g$ can be simplified as follows,

$$\begin{aligned}\nabla\Delta g &= \nabla\left(\sum_{j=1}^{d}\frac{\partial g}{\partial x_j^2}\right)\\ &= \left[\sum_{j=1}^{d}\frac{\partial}{\partial x_1}\frac{\partial g}{\partial x_j^2}, \sum_{j=1}^{d}\frac{\partial}{\partial x_2}\frac{\partial g}{\partial x_j^2}, \cdots, \sum_{j=1}^{d}\frac{\partial}{\partial x_d}\frac{\partial g}{\partial x_j^2},\right]\\ &= \left[\sum_{j=1}^{d}\frac{\partial^2 g}{\partial x_i\partial x_j^2},\right]_{i=1}^{d}\end{aligned} \tag{62}$$

Since (61) is equal to (62) it implies that

$$\Delta\nabla g = \nabla\Delta g.$$

□

Corollary 2. *For all vector valued function $f : \mathbb{R}^d \to \mathbb{R}^d$ functions the following holds,*

$$\nabla(-\Delta)^{-1}\mathrm{div_x}(f) = (-\Delta)^{-1}\mathrm{div_x}(\nabla f). \tag{63}$$

Proof. We know from Lemma 17 that for a vector valued function $f : \mathbb{R}^d \to \mathbb{R}^d$ that we have

$$\nabla\mathrm{div_x}(f) = \mathrm{div_x}(\nabla f).$$

Now, using for a fact that any function g can be written as, $g = (-\Delta)(-\Delta)^{-1}g$ we get,

$$\begin{aligned}&\nabla\mathrm{div_x}(f) = \mathrm{div_x}(\nabla f)\\ \implies &\nabla(-\Delta)(-\Delta)^{-1}\mathrm{div_x}(f) = \mathrm{div_x}(\nabla f)\\ \overset{(i)}{\implies} &(-\Delta)\nabla(-\Delta)^{-1}\mathrm{div_x}(f) = \mathrm{div_x}(\nabla f)\\ \implies &\nabla(-\Delta)^{-1}\mathrm{div_x}(f) = (-\Delta)^{-1}\mathrm{div_x}(\nabla f)\end{aligned}$$

where (i) follows from Lemma 18, i.e., for any function $g \in C^3$, we have, $\nabla\Delta g = \Delta\nabla g$. □

E.2. Some properties of Sub-Matrices

Lemma 19. *Given matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times d}$ if we have $A \preceq B$ then for any set of indices $U \subseteq \{1, 2, \cdots d\}$ where $|U| = n \leq d$ then for all $y \in \mathbb{R}^n$ we have $y^T A_U y \leq y^T B_U y$. where $A_U = A_{i,j}$ for all $i, j \in U$. Similarly if if we have $A \succeq B$ for all $y \in \mathbb{R}^n$ we have, $y^T A_U y \geq y^T B_U y$.*

Proof. We will show that $A \preceq B \implies A_U \preceq B_U$. The proof for $A \succeq B \implies A_U \succeq B_U$ will follow similarly.

Without loss of generality we can assume that $U = \{1, 2, \cdots n\}$ and a set $V = \{n, \cdots d\}$, where $n \leq d$. Since $A \preceq B$ we know that there exists $x \in \mathbb{R}^d$ we have $x^T Ax \leq x^T Bx$.

For all $y \in \mathbb{R}^d$ define $x := (y, \mathbf{0}_{d-n})$, and let $A_{U,V} = A_{i,j}$ be $i \in U$ and $j \in V$

$$\begin{aligned}
&\begin{bmatrix} y & \mathbf{0} \end{bmatrix}^T \begin{bmatrix} A_U & A_{U,V} \\ A_{V,U} & A_V \end{bmatrix} \begin{bmatrix} y & \mathbf{0} \end{bmatrix}^T \leq \begin{bmatrix} y & \mathbf{0} \end{bmatrix}^T \begin{bmatrix} B_U & B_{U,V} \\ B_{V,U} & B_V \end{bmatrix} \begin{bmatrix} y & \mathbf{0} \end{bmatrix}^T \\
\implies &\begin{bmatrix} y & \mathbf{0} \end{bmatrix}^T \begin{bmatrix} B_U - A_U & B_{U,V} - A_{U,V} \\ B_{V,U} - A_{V,U} & B_V - A_V \end{bmatrix} \begin{bmatrix} y & \mathbf{0} \end{bmatrix}^T \geq 0 \\
\implies &y^T (B_U - A_U) y \geq 0
\end{aligned}$$

Since we have for all $y \in \mathbb{R}^n$ we have $y^T(B_U - A_U)y \geq 0$, therefore this implies that $A_U \preceq B_U$. □