Approximation of Maximally Monotone Operators : A Graph Convergence Perspective

Takashi Furuya
Doshisha University, RIKEN AIP
tfuruya@mail.doshisha.ac.jp

Yury Korolev
Department of Mathematical Sciences
University of Bath, UK
ymk30@bath.ac.uk

Takaharu Yaguchi
Kyushu University, RIKEN AIP
yaguchi@imi.kyushu-u.ac.jp

Abstract

Operator learning has been highly successful for continuous mappings between infinite-dimensional spaces, such as PDE solution operators. However, many operators of interest—including differential operators—are discontinuous or set-valued, and lie outside classical approximation frameworks. We propose a paradigm shift by formulating approximation via graph convergence (Painlevé–Kuratowski convergence), which is well-suited for closed operators. We show that uniform and L^p approximation are fundamentally inadequate in this setting. Focusing on maximally monotone operators, we prove that any such operator can be approximated in the sense of local graph convergence by continuous encoder–decoder architectures, and further construct structure-preserving approximations that retain maximal monotonicity via resolvent-based parameterizations.

1 Introduction

Operator learning has emerged as a powerful paradigm for approximating mappings between infinite-dimensional spaces, with typical applications in partial differential equations (PDEs) or inverse problems [14, 34, 38, 41]. Most existing approaches focus on approximating solution operators of PDEs, mapping e.g. boundary conditions or coefficients to corresponding solutions. In contrast, a central perspective of this work is to learn the underlying differential operator itself, which defines the governing equation. Such methods constitute another important class of applications of machine learning to science [13, 21, 29, 37, 44]. This viewpoint is fundamentally different, as differential operators are typically discontinuous (unbounded) and may even be set-valued [1, 47], and thus fall outside classical approximation frameworks in operator learning. However, differential operators are naturally characterized as *closed operators*, i.e. those with a closed graph.

In this work we therefore focus on closed operators as targets. This class includes not only differential operators, but also many important operators arising in applications. For instance, inverses of compact operators and subdifferential operators of convex functionals fall into this class. Subdifferentials play a central role in variational problems [8, 45], while inverse operators are the subject of rich inverse problems and regularization theory [11, 25, 46].

A key difficulty is that classical approximation frameworks in operator learning rely on continuity or L^p-type structures, which fail to capture the behavior of closed operators. As a result, such operators cannot, in general, be approximated under local uniform or L^p convergence. This raises a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

fundamental question:

How should one formulate approximation when the target operator is closed?

To address the lack of continuity, we advocate changing the topology of approximation. A natural candidate is *graph convergence* (also known as Painlevé–Kuratowski convergence [2, Sec. 2.2]), which is well-established in functional and variational analysis [45].

The key advantage of graph convergence is that it ensures stability of the learned input–output relation under limiting operations, in a manner consistent with the notion of closed operators. In particular, any accumulation point of approximate input–output pairs remains valid for the limiting operator. This property is essential in applications where solutions are defined through limits, such as PDEs. Moreover, graph convergence naturally accommodates set-valued operators and situations where the domains of the approximating operators and the target operator differ. In contrast, classical notions such as uniform or L^p convergence are not well-defined in these settings. These features are particularly important in ill-posed inverse problems.

However, the class of closed operators is too broad to admit a meaningful universal approximation theory under graph convergence and we provide simple examples of closed operators that cannot be approximated by continuous operator models, not only in the uniform or L^2 sense, but also in the sense of graph convergence (Examples 2 to 4). This reveals a fundamental structural barrier: without additional structure, even graph convergence fails for closed operators.

In this work, we therefore consider maximally monotone operators [16], which provide a natural and practically relevant subclass of closed operators where meaningful approximation becomes possible. The most prominent example is the subdifferential of a convex functional, which includes many practically relevant differential operators such as the p-Laplace operator (which corresponds to the p-Dirichlet energy $J(u) = \frac{1}{p}\int \|\nabla u(x)\|^p \, dx$). More examples are given in Example 5, see also [10, 17].

A key advantage of this class is that it admits well-behaved resolvent mappings, which are Lipschitz continuous and defined on the entire space. This enables stable approximation via Yosida regularization [8, 45] and provides a natural bridge to continuous operator models. Moreover, maximal monotonicity is preserved under graph convergence [2], making it fully consistent with our approximation framework. In contrast, as shown in Example 4, there exist closed and monotone but not maximally monotone operators that cannot be approximated by any sequence of continuous operators even in graph convergence. Thus, monotonicity alone is not sufficient for a general graph-approximation result. This observation, together with the canonical single-valued regularizations available for maximally monotone operators, motivates our focus on maximal monotonicity as a natural structural assumption for the approximation theory developed in this work. Finally, even within the class of maximally monotone operators, approximation in uniform or L^p convergence is generally impossible (see Example 3). This further justifies graph convergence as the appropriate notion of approximation in our setting.

1.1 Related works

A large body of recent work establishes approximation guarantees under the assumption that the target operator is continuous or belongs to L^p, typically proving convergence in the local uniform topology [26, 32, 33, 35] or in L^p-type norms [12, 31, 36]. In contrast, our setting focuses on maximally monotone operators, which may be discontinuous, set-valued, and not belonging to a Bochner space. As a result, classical universality results in operator learning are not directly applicable in this setting.

Learning monotone or firmly non-expansive structures has also been considered in the context of inverse problems. [43] learns neural networks representing resolvents of maximally monotone operators, while [15] proposes a data-driven approach for constructing firmly non-expansive operators. [9] learns monotone neural networks via a penalization-based loss. However, these works do not address approximation of maximally monotone operators from the viewpoint of graph convergence.

Graph convergence of maximally monotone operators has been extensively studied in variational analysis and nonlinear PDEs; see, e.g., [2, 22]. In particular, [2] studies the preservation of maximal monotonicity under graph convergence. In contrast, our goal is not to deduce maximal monotonicity of approximating operators from graph convergence, but rather to construct approximations of a target maximally monotone operator using continuous encoder–decoder models, which are canonical

architectures in operator learning [26, 34]. Importantly, our approach enforces maximal monotonicity of the approximants by construction.

Unlike classical universality results based on global uniform convergence, the graph convergence framework considered here is inherently local. More precisely, the approximating encoder–decoder architectures depend on the compact set on which the graph approximation is evaluated.

Another related approach is the Input Convex Neural Network (ICNN) framework [3], which approximates convex functions and thereby induces monotone operators through their subdifferentials. In contrast, our framework applies to general maximally monotone operators, which need not be subdifferentials of convex functions and may therefore cover a broader class of set-valued operators.

1.2 Contribution

A summary of our contributions is as follows :

(1) **A paradigm shift in operator approximation.** We show, through explicit counterexamples, that classical approximation frameworks in operator learning – based on uniform or L^p convergence – are fundamentally inadequate for closed operators (Examples 2 and 3). These results indicate that the limitation lies not in the expressive power of models, but rather in the topology used to measure approximation. This suggests that the notion of approximation itself must be reconsidered.

(2) **Graph convergence as the appropriate topology.** We identify graph convergence as a natural notion of approximation for closed operators, since graph convergence measures discrepancies directly at the level of graphs rather than operator values. More precisely, we introduce a local graph distance (Definition 5) that quantitatively measures discrepancies between closed graphs on compact subsets. We further prove that convergence of this local graph distance to zero on every compact subset is equivalent to graph convergence (Proposition 3).

(3) **Approximation of maximally monotone operators in local graph convergence.** We show that monotonicity and closedness alone are insufficient for approximability even in graph convergence (Example 4). In contrast, within the class of maximally monotone operators, although approximation in uniform or L^p convergence still fails in general, approximation in the sense of local graph convergence becomes possible. In particular, we prove that any maximally monotone operator can be approximated by continuous encoder–decoder architectures (Theorem 1), while maximal monotonicity can also be preserved in the approximation process (Corollary 1).

We emphasize that the novelty of our work does not lie in the individual classical ingredients from maximal monotone operator theory, such as resolvents and Yosida approximations, themselves. Rather, our contribution is to combine these tools with recent approximation results in operator learning in order to develop an approximation theory for an unknown maximally monotone, and hence possibly set-valued, operator itself. In particular, we establish graph-convergence universality: the original set-valued operator can be approximated, in the graph sense, by single-valued encoder–decoder architectures arising from the operator-learning framework. Thus, our results bridge the classical theory of maximally monotone operators with operator-learning approximation theory, with convergence formulated directly at the level of the graph of the unknown operator.

1.3 Notation

Let H be a Hilbert space. We write $A : D(A) \subset H \to H$ to denote an operator A defined on the domain $D(A)$ with values in H. When we write $A : H \to H$, we mean that $D(A) = H$. We denote by $R(A)$ the range of A. If A is a set-valued operator we write $A : D(A) \subset H \to 2^H$. In this case, we define the graph of A by $\mathrm{gph}(A) := \{(u, v) \in H \times H \mid u \in D(A),\ v \in A(u)\}$. Finally, we denote the Euclidean norm on $\mathbb{R}^n$ by $|\cdot|$.

2 Non-Approximability of Closed Operators

In this section we define closed operators and several notions of convergence, and highlight the difficulties in establishing universal approximation results for closed operators in these topologies.

2.1 Closed operator

Definition 1 (Closed operator)**.** *We say that $A : D(A) \subset H \to 2^H$ is* closed *if for any sequences*

$$x_m \in D(A), \quad x_m \to x \text{ in } H, \qquad y_m \in A(x_m), \quad y_m \to y \text{ in } H,$$

it follows that $x \in D(A)$ and $y \in A(x)$. In other words, $\mathrm{gph}(A)$ *is closed in $H \times H$.*

Example 1. *Closed operators include differential operators and, more generally, inverses of compact operators, which naturally arise in scientific computing and inverse problems.*

(a) The differential operator $A : H^1(0,1) \subset L^2(0,1) \to L^2(0,1)$, $A(u)(x) = \frac{d}{dx}u(x)$.

(b) Laplacian operator $\Delta : H^2(\mathbb{R}^d) \subset L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$.

(c) The inverse of compact operator $A : R(K) \subset H \to 2^H$, $A = K^{-1}$ where $K : H \to H$ is compact.

2.2 Locally uniform convergence

Definition 2 (Locally uniform convergence)**.** *Let $A : D(A) \subset H \to H$ and $A_m : D(A_m) \subset H \to H$. We say that A_m converges to A in locally uniform convergence and write*

$$A_m \xrightarrow{\text{loc. unif.}} A$$

if for any compact set $K \subset H$ with $K \cap D(A) \cap D(A_m) \neq \emptyset$ for all m, it holds that

$$\sup_{x \in K \cap D(A) \cap D(A_m)} \|A_m(x) - A(x)\|_H \to 0.$$

Closed operators are, in general, not continuous and therefore cannot be obtained as a limit of continuous operators under locally uniform convergence (recall that a uniform limit of continuous operators is itself continuous). We illustrate this with the following simple example

Example 2. *Define the differential operator $A : H^1(0,1) \subset L^2(0,1) \to L^2(0,1)$ by*

$$A(u)(x) := \frac{d}{dx}u(x).$$

Then, A is a closed operator (Example 1 (a)). Define the compact set

$$K := \{0\} \cup \{v_n : n \in \mathbb{N}\} \subset D(A), \qquad v_n(x) := \frac{1}{n}\sin(n\pi x) \in H^1(0,1).$$

Then, for any $\varepsilon \in (0,1)$, there is no continuous operator $A_\varepsilon : D(A_\varepsilon) \subset L^2(0,1) \to L^2(0,1)$ with $K \subset D(A_\varepsilon)$ such that

$$\sup_{u \in K} \|A_\varepsilon(u) - A(u)\|_{L^2(0,1)} \leq \varepsilon.$$

The proof can be found in Appendix A.1.

2.3 L^p convergence

Definition 3 (L^p convergence)**.** *Let π be a probability measure on H. Let $1 \leq p < \infty$. Let $A : D(A) \subset H \to H$ and $A_m : D(A_m) \subset H \to H$ such that* $\mathrm{supp}(\pi) \subset D(A)$ *and* $\mathrm{supp}(\pi) \subset D(A_m)$ *for all m. We say that A_m converges to A in $L^p(\pi; H)$ and write*

$$A_m \xrightarrow{L^p(\pi;H)} A \quad \text{if} \quad \int \|A_m(u) - A(u)\|_H^p \, d\pi(u) \to 0.$$

Proposition 1. *Let π be a probability measure on H such that* $\mathrm{supp}(\pi)$ *is compact. Assume that* $\mathrm{supp}(\pi) \subset D(A)$ *and* $\mathrm{supp}(\pi) \subset D(A_m)$ *for all m. Then it holds that*

$$A_m \xrightarrow{\text{loc. unif.}} A \implies A_m \xrightarrow{L^p(\pi;H)} A.$$

The proof can be found in Appendix A.2.

The next example shows that, although $L^p(\pi; H)$-convergence is weaker than locally uniform convergence, in general a closed operator cannot be approximated by continuous ones in the $L^p(\pi; H)$ sense.

Example 3. *Define the differential operator* $A : H^1_{\mathrm{per}}(0,1) \subset L^2_{\mathrm{per}}(0,1) \to L^2_{\mathrm{per}}(0,1)$, *by*

$$A(u)(x) := \frac{d}{dx}u(x).$$

Then, A is a closed operator. Moreover, A is maximally monotone (See Definition 6). Define a probability measure

$$\pi := \sum_{n=2}^{\infty} w_n \, \delta_{u_n}, \quad w_n := \frac{c}{n^{1+p/2}}, \quad u_n(x) := \frac{1}{\sqrt{n}} \sin(2\pi n x)$$

where $c > 0$ is chosen so that $\sum_{n=2}^{\infty} w_n = 1$. Then, for any $\varepsilon \in (0,1)$, there is no continuous $A_\varepsilon : D(A_\varepsilon) \subset L^2_{\mathrm{per}}(0,1) \to L^2_{\mathrm{per}}(0,1)$ with $\mathrm{supp}(\pi) \subset D(A_\varepsilon)$ such that

$$\int \|A_\varepsilon(u) - A(u)\|^p_{L^2(0,1)} d\pi(u) \le \varepsilon.$$

The proof can be found in Appendix A.3.

Example 3 does not imply that an unbounded operator cannot be approximated in an L^p topology for every input measure. Rather, it shows that there exist probability measures for which such an approximation is impossible. In particular, the obstruction depends on the interplay between the growth of the operator and the tail or regularity properties of the input distribution. For sufficiently regular input measures, including suitable Gaussian measures, unbounded linear operators may admit approximation in an L^2 topology; see, e.g., [24].

2.4 Graph convergence

Now we introduce the graph (Painlevé-Kuratowski) convergence, see [2, Sec. 2.2].

Definition 4 (Graph convergence)**.** *Let $A : D(A) \subset H \to 2^H$ and $A_m : D(A_m) \subset H \to 2^H$. We say that A_m converges to A in the sense of* graph convergence *and write*

$$A_m \xrightarrow{g} A$$

if $\begin{cases} \text{(outer limit condition) } (x_m, y_m) \in \mathrm{gph}(A_m),\ (x_m, y_m) \to (x, y) \ \Rightarrow\ (x, y) \in \mathrm{gph}(A) \\ \text{(inner limit condition) } \forall (x, y) \in \mathrm{gph}(A)\ \exists (x_m, y_m) \in \mathrm{gph}(A_m) \text{ such that } (x_m, y_m) \to (x, y). \end{cases}$

Equivalently, the outer limit condition may be stated for every subsequence $(m_k)_{k\in\mathbb{N}}$.

Graph convergence intuitively means that $\mathrm{gph}(A_m)$ converge to $\mathrm{gph}(A)$ as subsets of $H \times H$. While the locally uniform and $L^p(\pi; H)$ convergences are defined for single-valued operators, graph convergence makes sense for set-valued operators. The following proposition shows that when operators are single-valued, the graph convergence is weaker than the locally uniform convergence.

Proposition 2. *Let $A : D(A) \subset H \to H$ and $A_m : D(A_m) \subset H \to H$ be single-valued operators. Assume that $D(A) = D(A_m)$, and A is closed. Then it holds that*

$$A_m \xrightarrow{\text{loc. unif.}} A \implies A_m \xrightarrow{g} A.$$

The proof can be found in Appendix A.4.

Just as with $L^p(\pi; H)$-convergence, even though graph convergence is weaker than locally uniform one, an arbitrary closed operator cannot be approximated by continuous ones in the graph sense.

Example 4. *Let $H = \mathbb{R}$. Define a set-valued operator $A : \mathbb{R} \to 2^{\mathbb{R}}$ by $D(A) = \mathbb{R}$ and*

$$A(x) := \begin{cases} \{0\}, & x < 0, \\ \{0, 1\}, & x = 0, \\ \{1\}, & x > 0. \end{cases}$$

Then A is closed. Moreover, A is monotone, but not maximally monotone (See Definition 6). In addition, there is no sequence of continuous operators $A_m : \mathbb{R} \to \mathbb{R}$ such that

$$A_m \xrightarrow{g} A.$$

The proof can be found in A.5.

Now we define a notion of distance that will play an important role in our approximation results.

Definition 5 (Local graph distance)**.** *For $A : D(A) \subset H \to 2^H$ and $B : D(B) \subset H \to 2^H$, define the* local graph distance *for any compact set $K \subset H \times H$*

$$d_{\mathrm{gph},K}(A, B) := \max\left\{h_K(\mathrm{gph}(B), \mathrm{gph}(A)), h_K(\mathrm{gph}(A), \mathrm{gph}(B))\right\},$$

where h_K is defined by for $X, Y \subset H \times H$

$$h_K(X, Y) := \begin{cases} \sup_{(x,y)\in X\cap K} \inf_{(u,v)\in Y} \|(x, y) - (u, v)\|_{H\times H}, & X \cap K \neq \emptyset, \\ 0, & X \cap K = \emptyset. \end{cases}$$

The reason for introducing the local graph distance is that graph convergence, as defined in Definition 4, is qualitative in nature. In contrast, the local graph distance provides a quantitative measure of discrepancies between graphs on compact subsets, which allows us to formulate quantitative approximation results.

In the finite-dimensional setting, [2, Remark 2.1] shows that graph convergence coincides with the ρ-Hausdorff convergence, which is induced by a genuine metric topology. In infinite-dimensional Hilbert spaces, however, extending such a metric characterization is nontrivial. For this reason, we introduce a weaker local graph distance, which is generally not a metric. Nevertheless, the following proposition shows that it still characterizes graph convergence equivalently.

Proposition 3. *Let $A : D(A) \subset H \to 2^H$ and $A_m : D(A_m) \subset H \to 2^H$ be closed. Then the following are equivalent :*

$$A_m \xrightarrow{g} A \quad \Longleftrightarrow \quad d_{\mathrm{gph},K}(A_m, A) \to 0 \text{ for every compact } K \subset H \times H.$$

We refer to Appendix A.6 for the proof.

3 Approximation of Maximally Monotone Operators

In the previous section, we discussed the difficulty of approximating closed operators in the sense of uniform, L^p, and even graph convergence. A natural solution is to impose additional structure on the target operator. In this section, we focus on maximally monotone operators and establish their universal approximation.

3.1 Maximally monotone operators

Definition 6 (Monotone and maximally monotone operators [16])**.** *We say that $A : D(A) \subset H \to 2^H$ is* monotone *if*

$$\langle y - v,\, x - u\rangle_H \geq 0 \quad \forall (x, y), (u, v) \in \mathrm{gph}(A).$$

We say that monotone operator $A : H \to 2^H$ is maximally monotone *if there exists no monotone operator $B : D(B) \subset H \to 2^H$ such that $\mathrm{gph}(B)$ properly contains $\mathrm{gph}(A)$.*

It is well known that any maximally monotone operator is closed (e.g., [8, Prop. 20.33]).

Proposition 4. *If $A : D(A) \subset H \to 2^H$ is maximally monotone, then A is closed.*

Example 5. *We give some examples of maximally monotone operators. Note that appropriate boundary conditions may be required to ensure maximal monotonicity.*

(a) Subdifferentials of convex functionals $H \to \mathbb{R} \cup \{+\infty\}$. For example, the p-Laplacian, $1 < p < +\infty$, $A : W^{1,p}(\Omega) \subset L^2(\Omega) \to L^2(\Omega)$ for a bounded open set $\Omega \subset \mathbb{R}^d$, defined as $A(u) = -\mathrm{div}(|\nabla u|^{p-2}\nabla u)$, which is a subgradient of the p-Dirichlet energy functional $J\colon L^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ given by

$$J(u) = \begin{cases} \frac{1}{p}\int_\Omega \|\nabla u\|^p \, dx, & u \in W^{1,p}(\Omega), \\ +\infty, & u \in L^2(\Omega) \setminus W^{1,p}(\Omega). \end{cases}$$

(b) The differentiation operator (the inverse Volterra operator [7]) on periodic functions $A : H^1_{\mathrm{per}}(0,1) \subset L^2_{\mathrm{per}}(0,1) \to L^2_{\mathrm{per}}(0,1)$, $A(u)(x) = \frac{d}{dx}u(x)$ *(which is not symmetric and therefore cannot be the subdifferential of a convex function).*

More examples can be found in Appendix B and in [10].

Remark 1. *We give several remarks on the approximability of maximally monotone operators with respect to the choice of convergence topology.*

(1) Example 5 (b) shows that the differential operator $A : H^1_{\mathrm{per}}(0,1) \subset L^2_{\mathrm{per}}(0,1) \to L^2_{\mathrm{per}}(0,1)$, $A(u)(x) := \frac{d}{dx}u(x)$ *is maximally monotone. On the other hand, Example 3 shows that* $A = \frac{d}{dx}$ *cannot be approximated by any sequence of continuous operators in* L^p*, which also rules out uniform convergence by Proposition 1. This demonstrates that, even under maximal monotonicity, approximation in* L^p *or uniform convergence may fail.*

(2) Example 4 shows that there exist closed and monotone but not maximally monotone operators that cannot be approximated by any sequence of continuous operators even in the graph sense. This indicates that maximal monotonicity is a sharp structural condition for approximability.

Definition 7 (Yosida approximation)**.** *The* Yosida approximation *of maximally monotone operator* $A : D(A) \subset H \to 2^H$ *is a family* $(A_\lambda)_{\lambda>0}$ *of single-valued operator* $: H \to H$ *defined by*

$$A_\lambda := \frac{I - J_\lambda}{\lambda}, \quad J_\lambda := (I + \lambda A)^{-1}.$$

By Minty's theorem [8, Thm. 21.1], the Yosida approximation $(A_\lambda)_{\lambda>0}$, and in particular the resolvent J_λ, is well-posed. Moreover, the Yosida approximation $(A_\lambda)_{\lambda>0}$ has the following properties (e.g., [8, Cor. 23.10, Cor. 23.46]).

Proposition 5. *Let* $A : D(A) \subset H \to 2^H$ *be maximally monotone.* $A_\lambda : H \to H$ *is* $\frac{1}{\lambda}$ *Lipschitz continuous and maximally monotone. Moreover, it holds that for* $x \in D(A)$, $\lim_{\lambda\to 0} A_\lambda(x) = A^0(x)$, *where* $A^0(x)$ *denotes the element of minimal norm in* $A(x)$.

Note that the above proposition states that, if A is a single-valued maximally monotone operator, then A^0 coincides with A. In particular, this implies that A_λ converges to A pointwise. On the other hand, the following proposition shows that the Yosida approximation of a (possibly set-valued) maximally monotone operator converges to A in the sense of graph convergence.

Proposition 6. *Let* $A : D(A) \subset H \to 2^H$ *be maximally monotone. Then it holds that*

$$A_\lambda \xrightarrow{g} A \quad (\lambda \to 0)$$

A proof of this statement in the finite-dimensional case $H = \mathbb{R}^n$ can be found in [45] but it holds for a general Hilbert space H, see Appendix A.7.

3.2 Encoder–decoder architectures

Definition 8 (Encoder–decoder approximation property)**.** *A Hilbert space* H *is said to have the encoder–decoder approximation property (EDAP) if there exist sequences of 1-Lipschitz continuous maps*

$$E_m : H \to \mathbb{R}^{w_m}, \qquad D_m : \mathbb{R}^{w_m} \to H,$$

with $E_m(0) = 0$, $D_m(0) = 0$, *and* $(w_m)_{m\in\mathbb{N}} \subset \mathbb{N}$ *with* $w_m \to \infty$, *such that for any compact set* $K \subset H$, *it holds that*

$$\lim_{m\to\infty} \sup_{u\in K} \|u - D_m \circ E_m(u)\|_H = 0.$$

Definition 9 (Universal approximator)**.** *A class* $\mathcal{F}$ *of continuous functions is called a universal approximator if, for any* $p \in \mathbb{N}$, *any compact set* $K \subset \mathbb{R}^p$, *and any continuous function* $\varphi : \mathbb{R}^p \to \mathbb{R}^p$, *there exists a sequence* $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ *such that*

$$\lim_{n\to\infty} \sup_{x\in K} |\varphi_n(x) - \varphi(x)| = 0. \tag{1}$$

Similarly, a class $\mathcal{F}^{\mathrm{Lip}}$ *of 1-Lipschitz continuous functions is called a 1-Lipschitz universal approximator if, for any* $p \in \mathbb{N}$, *any compact set* $K \subset \mathbb{R}^p$, *and any 1-Lipschitz continuous function* $\varphi : \mathbb{R}^p \to \mathbb{R}^p$, *there exists a sequence* $(\varphi_n)_{n\in\mathbb{N}} \subset \mathcal{F}^{\mathrm{Lip}}$ *such that such that (1) holds.*

Definition 10 (Encoder–decoder architectures)**.** *Let H be a Hilbert space satisfying EDAP with encoder–decoder pairs (E_m, D_m). Let $\mathcal{F}$ (resp. $\mathcal{F}^{\mathrm{Lip}}$) be a (1-Lipschitz) universal approximator. We call operators of the form*

$$D_m \circ \varphi \circ E_m : H \to H, \qquad \varphi \in \mathcal{F} \text{ (resp. } \mathcal{F}^{\mathrm{Lip}}),$$

(1-Lipschitz) encoder–decoder architectures.

The encoder–decoder architectures $D_m \circ \varphi \circ E_m$ constitute a canonical class of operator models; see, e.g., [26, 34]. For example, classes such as DeepONet and neural operators can be interpreted within this framework.

A typical example for encoder-decoder pair (E_m, D_m) is as follows: let H be a Hilbert space with a complete orthonormal basis $(e_i)_{i\in\mathbb{N}}$. Define E_m as the orthogonal projection onto the subspace $\mathrm{span}\{e_i : i \le m\}$, and let D_m be its adjoint operator. Then both E_m and D_m are 1-Lipschitz continuous.

The class $\mathcal{F}$ can be taken as a family of neural networks that are universal approximators for continuous functions on Euclidean space [23, 27, 28]. Similarly, the class $\mathcal{F}^{\mathrm{Lip}}$ can be chosen as neural networks that preserve 1-Lipschitz continuity, which have been studied in recent works such as [4, 39, 40, 42].

Recent work [26, Thm. 4.4] shows that encoder-decoder architecture ensure the locally uniform convergence to any continuous operator.

Proposition 7. *Let $\mathcal{F}$ be universal approximator, and let H have the EDAP with E_m and D_m. Let $A : D(A) \subset H \to H$ be a continuous operator. Then, there exist a sequence $(\varphi_m)_{m\in\mathbb{N}} \subset \mathcal{F}$ such that*

$$D_m \circ \varphi_m \circ E_m \xrightarrow{\text{loc. unif.}} A, \quad (m \to \infty). \tag{2}$$

Moreover, if $A : D(A) \subset H \to H$ is 1-Lipschitz continuous, and $\mathcal{F}^{\mathrm{Lip}}$ is 1-Lipschitz universal approximator, then, there exist a sequence $(\varphi_m)_{m\in\mathbb{N}} \subset \mathcal{F}^{\mathrm{Lip}}$ such that encoder–decoder architectures $D_m \circ \varphi_m \circ E_m$ is 1-Lipschitz continuous, and (2) holds.

3.3 Approximation result

We first establish the following main result that

Theorem 1. *Let $\mathcal{F}$ be a universal approximators, and assume that H has the EDAP with encoder and decoder mappings E_m and D_m. Let $A : D(A) \subset H \to 2^H$ be maximally monotone. Then, for any compact set $K \subset H \times H$, there exists a sequence $(\varphi_m)_{m\in\mathbb{N}} \subset \mathcal{F}$ such that as $m \to \infty$*

$$d_{\mathrm{gph},K}(A_m, A) \to 0, \quad A_m := D_m \circ \varphi_m \circ E_m.$$

This theorem states that any maximally monotone operator can be approximated by continuous encoder–decoder architectures in the sense of local graph convergence. The proof is based on using the Yosida approximation A_λ (see Proposition 6), which is approximated by encoder–decoder architectures via Proposition 7. We refer to Appendix A.8 for details. Moreover, using quantitative approximation results for neural networks applied to the Yosida approximation, we can obtain a quantitative error estimates. See Appendix C for details.

Remark 2. *The approximation result in Theorem 1 is local in the following sense: for each compact set $K \subset H \times H$, the approximating sequence may depend on K. Thus, in a general infinite-dimensional Hilbert space, Theorem 1 does not assert the existence of a single sequence of encoder–decoder architectures that converges graphically to A on the whole space.*

If, in addition, $H \times H$ is hemicompact, these local approximations can be combined into a single graph-convergent sequence. Indeed, let

$$K_1 \subset K_2 \subset \cdots \subset H \times H$$

be a sequence of compact sets such that every compact subset of $H \times H$ is contained in some K_n. For each n, Theorem 1 yields a sequence $(A_m^{(n)})_{m\in\mathbb{N}}$ of encoder–decoder architectures such that

$$d_{\mathrm{gph},K_n}(A_m^{(n)}, A) \to 0 \qquad (m \to \infty).$$

Hence, for each n, we may choose $m_n \geq n$ sufficiently large so that

$$d_{\mathrm{gph},K_n}(A^{(n)}_{m_n}, A) < \frac{1}{n}.$$

Setting

$$\widetilde{A}_n := A^{(n)}_{m_n},$$

we obtain, for every compact set $K \subset H \times H$,

$$d_{\mathrm{gph},K}(\widetilde{A}_n, A) \to 0.$$

Indeed, by hemicompactness there exists N such that $K \subset K_N$, and hence $K \subset K_n$ for every $n \geq N$. Therefore,

$$d_{\mathrm{gph},K}(\widetilde{A}_n, A) \leq d_{\mathrm{gph},K_n}(\widetilde{A}_n, A) < \frac{1}{n}, \qquad n \geq N.$$

By Proposition 3, it follows that

$$\widetilde{A}_n \xrightarrow{g} A.$$

In particular, this applies when H is finite-dimensional.

Moreover, for maximally monotone operators, graph convergence is classically equivalent to convergence of the associated resolvents. Consequently, whenever the above construction yields $\widetilde{A}_n \xrightarrow{g} A$ and the approximants $\widetilde{A}_n$ are maximally monotone, for every fixed $\lambda > 0$,

$$(I + \lambda \widetilde{A}_n)^{-1} x \longrightarrow (I + \lambda A)^{-1} x \qquad \text{for every } x \in H;$$

see, e.g., [6, Proposition 3.2]. Thus, in particular, a single graph convergent sequence of maximally monotone approximants also yields convergence of the corresponding resolvents. In contrast, the compact-dependent approximation provided by Theorem 1 alone does not yield such a conclusion in a general infinite-dimensional Hilbert space.

Finally, we emphasize that graph-based approximation and resolvent-based approximation need not correspond to different information settings. Indeed, a graph sample (x, y) with $y \in A(x)$ satisfies

$$J_\lambda(x + \lambda y) = x, \qquad J_\lambda := (I + \lambda A)^{-1},$$

and therefore directly provides the input–output pair $(x + \lambda y, x)$ for learning the resolvent J_λ. The distinction is instead one of approximation target. In the present work, the possibly set-valued operator A itself is the object of approximation, and consistency is formulated directly in terms of its graph. Resolvent-based approximation, by contrast, takes the single-valued map J_λ for a fixed $\lambda > 0$ as the approximation target. Our graph-based formulation should therefore not be viewed as a replacement for resolvent-based approaches, but rather as an approximation framework in which the original operator itself is the target object.

While the above encoder–decoder architectures do not, in general, preserve maximal monotonicity, we can establish the following structure-preserving approximation result inspired by the resolvent formulation of the Yosida approximation.

Corollary 1. *Let $\mathcal{F}^{\mathrm{Lip}}$ be a class of 1-Lipschitz universal approximators, and assume that H has the EDAP with E_m and D_m. Let $A : D(A) \subset H \to 2^H$ be maximally monotone. Then, for any compact set $K \subset H \times H$, there exist a sequence $(\lambda_m)_{m\in\mathbb{N}} \subset \mathbb{R}_+$ with $\lambda_m \to 0$ and a sequences $(\psi_m)_{m\in\mathbb{N}} \subset \mathcal{F}^{\mathrm{Lip}}$ such that*

$$d_{\mathrm{gph},K}(B_m, A) \to 0, \quad B_m := \frac{I - D_m \circ \psi_m \circ E_m}{2\lambda_m}.$$

Here, $D_m \circ \psi_m \circ E_m$ is a 1-Lipschitz encoder–decoder architecture, and $B_m : H \to H$ is maximally monotone.

The key idea is based on the resolvent representation of the Yosida approximation $A_\lambda = \frac{I-(2J_\lambda - I)}{2\lambda}$. Here, $2J_\lambda - I$ is non-expansive (i.e., 1-Lipschitz). We approximate this mapping by 1-Lipschitz encoder–decoder architectures, which yields a sequence of maximally monotone operators by construction. We refer to Appendix A.9 for details.

4 Conclusion

We studied the universal approximation of maximally monotone operators from the viewpoint of graph convergence, rather than classical uniform or L^p convergence.

The limitations of our work and possible future directions are summarized as follows:

(a) Theorem 1 states that any maximally monotone operator can be approximated by continuous encoder–decoder architectures in the sense of local graph convergence. Note, however, that the encoder–decoder architectures constructed in the theorem depend on the compact set. Therefore, the theorem does not guarantee global graph convergence; see Proposition 3. This is one limitation of our approach. Under additional hemicompactness, however, the local approximations can be diagonalized into a single graph-convergent sequence, which also yields convergence of the associated resolvents; see Remark 2.

(b) Although Corollary 2 in Appendix C provides an upper bound on the local graph distance between the target operator and the encoder–decoder architecture, the bound is not fully explicit and is not uniform over compact sets (see Remark 3). Deriving sharper and more explicit quantitative estimates remains an important direction for future work.

(c) Numerical experiments illustrating our graph-based approximation framework are presented in Appendix D. A practical limitation of the graph-based loss used in these experiments is its computational cost. While standard losses such as ℓ^2 and soft ℓ^∞ have linear complexity $\mathcal{O}(N)$ in the number of samples, the graph loss involves pairwise comparisons between samples, resulting in quadratic complexity $\mathcal{O}(N^2)$. This computational overhead may limit scalability to large datasets. Developing computationally efficient approximations of graph-based distances is therefore an important direction for future work.

Acknowledgments

TF and TY are supported by JSPS KAKENHI Grant Number JP24K16949, 25H01453, 25K15148, JST CREST JPMJCR24Q5, JST ASPIRE JPMJAP2329. TF and YK acknowledge the support of the Royal Society (International Collaboration Award ICA\R2\252033). This work is also supported by the Horizon Europe, MSCA-SE project 101131557 (REMODEL.)

References

[1] Vincent Acary and Bernard Brogliato. *Numerical methods for nonsmooth dynamical systems: Applications in mechanics and electronics*. Lecture Notes in Applied and Computational Mechanics. Springer, Berlin, Germany, 2008.

[2] Samir Adly, Hédy Attouch, and Ralph Tyrrell Rockafellar. Preservation or not of the maximally monotone property by graph-convergence. 2022.

[3] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International conference on machine learning*, pages 146–155. PMLR, 2017.

[4] Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International conference on machine learning*, pages 291–301. PMLR, 2019.

[5] Gunnar Aronsson, Michael Crandall, and Petri Juutinen. A tour of the theory of absolutely minimizing functions. *Bulletin of the American mathematical society*, 41(4):439–505, 2004.

[6] H Attouch, JB Baillon, and M Théra. Variational sum of monotone operators. *J. Convex Anal*, 1(1):1–29, 1994.

[7] Heinz H Bauschke, Xianfu Wang, and Liangjin Yao. Examples of discontinuous maximal monotone linear operators and the solution to a recent problem posed by bf svaiter. *Journal of Mathematical Analysis and Applications*, 370(1):224–241, 2010.

[8] HH Bauschke. Convex analysis and monotone operator theory in hilbert spaces, 2011.

[9] Younes Belkouchi, Jean-Christophe Pesquet, Audrey Repetti, and Hugues Talbot. Learning truly monotone operators with applications to nonlinear inverse problems. *SIAM Journal on Imaging Sciences*, 18(1):735–764, 2025.

[10] Irene Benedetti, Luisa Malaguti, and Manuel DP Monteiro Marques. Differential equations with maximal monotone operators. *Journal of Mathematical Analysis and Applications*, 539(1):128484, 2024.

[11] Martin Benning and Martin Burger. Modern regularization methods for inverse problems. *Acta numerica*, 27:1–111, 2018.

[12] Kaushik Bhattacharya, Bamdad Hosseini, Nikola B Kovachki, and Andrew M Stuart. Model reduction and neural networks for parametric pdes. *The SMAI journal of computational mathematics*, 7:121–157, 2021.

[13] Gert-Jan Both, Subham Choudhury, Pierre Sens, and Remy Kusters. DeepMoD: Deep learning for model discovery in noisy data. *J. Comput. Phys.*, 428(109985):109985, March 2021.

[14] Nicolas Boullé and Alex Townsend. A mathematical guide to operator learning. In *Handbook of Numerical Analysis*, volume 25, pages 83–125. Elsevier, 2024.

[15] Kristian Bredies, Jonathan Chirinos-Rodriguez, and Emanuele Naldi. Learning firmly nonexpansive operators. *arXiv:2407.14156*, 2024.

[16] Haim Brezis. *Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, volume 5. Elsevier, 1973.

[17] Leon Bungert and Martin Burger. Asymptotic profiles of nonlinear homogeneous evolution equations of gradient flow type. *Journal of Evolution Equations*, 20(3):1061–1092, 2020.

[18] Leon Bungert and Martin Burger. Gradient flows and nonlinear power methods for the computation of nonlinear eigenfunctions. In *Handbook of numerical analysis*, volume 23, pages 427–465. Elsevier, 2022.

[19] Leon Bungert and Yury Korolev. Eigenvalue problems in L^∞: optimality conditions, duality, and relations with optimal transport. *Communications of the American Mathematical Society*, 2(08):345–373, 2022.

[20] Leon Bungert and Yury Korolev. Introduction to nonlinear spectral analysis. *arXiv:2506.08754*, 2025.

[21] Zhao Chen, Yang Liu, and Hao Sun. Physics-informed learning of governing equations from scarce data. *Nat. Commun.*, 12(1):6136, October 2021.

[22] Valeria Chiado'Piat, Gianni Dal Maso, and Anneliese Defranceschi. G-convergence of monotone operators. In *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, volume 7, pages 123–160. Elsevier, 1990.

[23] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[24] Maarten V de Hoop, Nikola B Kovachki, Nicholas H Nelsen, and Andrew M Stuart. Convergence rates for learning linear operators from noisy data. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):480–513, 2023.

[25] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[26] Janek Gödeke and Pascal Fernsel. New universal operator approximation theorem for encoder-decoder architectures (preprint). *arXiv:2503.24092*, 2025.

[27] Boris Hanin and Mark Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv:1710.11278*, 2017.

[28] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[29] Xianglong Hou, Xinquan Huang, and Paris Perdikaris. CFO: Learning continuous-time PDE dynamics via flow-matched neural operators. In *The Fourteenth International Conference on Learning Representations*, 2025.

[30] Petri Juutinen, Peter Lindqvist, and Juan J Manfredi. *The infinity Laplacian: examples and observations*. Institut Mittag-Leffler, 1999.

[31] Yury Korolev. Two-layer neural networks with values in a banach space. *SIAM Journal on Mathematical Analysis*, 54(6):6358–6389, 2022.

[32] Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021.

[33] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

[34] Nikola B Kovachki, Samuel Lanthaler, and Andrew M Stuart. Operator learning: Algorithms and analysis. *Handbook of Numerical Analysis*, 25:419–467, 2024.

[35] Samuel Lanthaler, Zongyi Li, and Andrew M Stuart. Nonlocality and nonlinearity implies universality in operator learning. *Constructive Approximation*, 62(2):261–303, 2025.

[36] Samuel Lanthaler, Siddhartha Mishra, and George E Karniadakis. Error estimates for deeponets: A deep learning framework in infinite dimensions. *Transactions of Mathematics and its Applications*, 6(1):tnac001, 2022.

[37] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. PDE-net: Learning PDEs from data. In *International Conference on Machine Learning*, pages 3208–3216. PMLR, July 2018.

[38] Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv:1910.03193*, 2019.

[39] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv:1802.05957*, 2018.

[40] Davide Murari, Takashi Furuya, and Carola-Bibiane Schönlieb. Approximation theory for 1-lipschitz resnets. *arXiv:2505.12003*, 2025.

[41] Nicholas H. Nelsen and Yunan Yang. Operator learning meets inverse problems: A probabilistic perspective. *arxiv:2508.20207*, 2025.

[42] Sebastian Neumayer, Alexis Goujon, Pakshal Bohra, and Michael Unser. Approximation of lipschitz functions using deep spline neural networks. *SIAM Journal on Mathematics of Data Science*, 5(2):306–322, 2023.

[43] Jean-Christophe Pesquet, Audrey Repetti, Matthieu Terris, and Yves Wiaux. Learning maximally monotone operators for image recovery. *SIAM Journal on Imaging Sciences*, 14(3):1206–1237, 2021.

[44] Maziar Raissi. Deep hidden physics models: deep learning of nonlinear partial differential equations. *J. Mach. Learn. Res.*, 19(1):932–955, 2018.

[45] R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

[46] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, and Frank Lenzen. *Variational methods in imaging*, volume 167. Springer.

[47] Yeh Wang. Finite element analysis of the duct flow of bingham plastic fluids: an application of the variational inequality. *Int. J. Numer. Methods Fluids*, 25(9):1025–1042, 1997.

[48] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94:103–114, 2017.

A Proofs

A.1 Proof of Example 2

Proof. Suppose, for contradiction, there is a continuous $A_\varepsilon : D(A_\varepsilon) \subset H \to H$ with $K \subset D(A_\varepsilon)$

$$\sup_{u\in K} \|A_\varepsilon(u) - A(u)\|_{L^2(0,1)} \le \varepsilon.$$

In particular,

$$\|A_\varepsilon(0)\|_{L^2(0,1)} = \|A_\varepsilon(0) - A(0)\|_{L^2(0,1)} \le \varepsilon.$$

We observe that since $\|Av_n\|_{L^2(0,1)} = \frac{\pi}{\sqrt{2}}$

$$\begin{aligned}\sup_{u\in K} \|A_\varepsilon(u) - A(u)\|_{L^2(0,1)} &\ge \|A_\varepsilon(v_n) - A(v_n)\|_{L^2(0,1)} \\ &\ge \|A(v_n)\|_{L^2(0,1)} - \|A_\varepsilon(v_n)\|_{L^2(0,1)} = \frac{\pi}{\sqrt{2}} - \|A_\varepsilon(v_n)\|_{L^2(0,1)}.\end{aligned}$$

Taking the limit as $n \to \infty$, using $v_n \to 0$ in $L^2(0,1)$, and using the continuity of A_ε

$$\varepsilon \ge \sup_{u\in K} \|A_\varepsilon(u) - A(u)\|_{L^2(0,1)} \ge \frac{\pi}{\sqrt{2}} - \|A_\varepsilon(0)\|_{L^2(0,1)} \ge \frac{\pi}{\sqrt{2}} - 1 > 0$$

This contradicts with the fact that $\varepsilon \in (0,1)$ can be arbitrary small. □

A.2 Proof of Proposition 1

Proof. Let $K := \mathrm{supp}(\pi)$. Since K is compact and $A_m \xrightarrow{\text{loc. unif.}} A$, we have

$$\sup_{u\in K} \|A_m(u) - A(u)\|_H \to 0.$$

Then

$$\int_H \|A_m(u) - A(u)\|_H^p \, d\pi(u) = \int_K \|A_m(u) - A(u)\|_H^p \, d\pi(u),$$

because π is supported on K. Hence

$$\int_H \|A_m(u) - A(u)\|_H^p \, d\pi(u) \le \left(\sup_{u\in K} \|A_m(u) - A(u)\|_H\right)^p \pi(K).$$

Since π is a probability measure and $\pi(K) = 1$, we obtain

$$\int_H \|A_m(u) - A(u)\|_H^p \, d\pi(u) \le \left(\sup_{u\in K} \|A_m(u) - A(u)\|_H\right)^p \to 0.$$

Therefore,

$$A_m \xrightarrow{L^p(\pi;H)} A.$$

□

A.3 Proof of Example 3

Proof. We first show that A is maximally monotone. Indeed, for any $u \in D(A)$

$$\langle Au, u\rangle_{L^2} = \int_0^1 u'(x)u(x)\,dx = \frac{1}{2}\int_0^1 (u(x)^2)'dx = \frac{1}{2}\{u(1)^2 - u(0)^2\} = 0.$$

Hence A is monotone.

We next show maximality by Minty's theorem. It suffices to prove that

$$\mathrm{Ran}(I + A) = H.$$

Let $f \in H$. We solve

$$u + u' = f, \qquad u(0) = u(1).$$

The general solution is

$$u(x) = e^{-x}\left(c + \int_0^x e^t f(t)\,dt\right),$$

where $c \in \mathbb{R}$ is a constant. The periodicity condition gives

$$c = e^{-1}\left(c + \int_0^1 e^t f(t)\,dt\right),$$

and hence

$$c = \frac{1}{e-1}\int_0^1 e^t f(t)\,dt.$$

Therefore there exists $u \in H^1_{\mathrm{per}}(0,1)$ such that

$$(I + A)u = f.$$

Thus

$$\mathrm{Ran}(I + A) = H.$$

By Minty's theorem, A is maximally monotone.

We finally prove the non-approximability statement. Suppose, for contradiction, that there exists a continuous operator $A_\varepsilon : D(A_\varepsilon) \subset L^2_{\mathrm{per}}(0,1) \to L^2_{\mathrm{per}}(0,1)$ with $\mathrm{supp}(\pi) \subset D(A_\varepsilon)$ such that

$$\int \|A_\varepsilon(u) - A(u)\|_H^p \, d\pi(u) \le \varepsilon.$$

Since $u_n \to 0$ in H, and since $0 \in \mathrm{supp}(\pi) \subset D(A_\varepsilon)$, the continuity of A_ε implies

$$A_\varepsilon(u_n) \to A_\varepsilon(0) \quad \text{in } H.$$

Hence there exist $M > 0$ and $N \in \mathbb{N}$ such that

$$\|A_\varepsilon(u_n)\|_H \le M \qquad (n \ge N).$$

On the other hand,

$$\|Au_n\|_H = \sqrt{2n}\pi.$$

Therefore, for sufficiently large n,

$$\|A_\varepsilon(u_n) - Au_n\|_H \ge \|Au_n\|_H - \|A_\varepsilon(u_n)\|_H \ge \frac{\sqrt{2n}\pi}{2}.$$

Consequently,

$$\begin{aligned}\int \|A_\varepsilon(u) - A(u)\|_H^p \, d\pi(u) &= \sum_{n=2}^{\infty} w_n \|A_\varepsilon(u_n) - A(u_n)\|_H^p \\ &\ge \sum_{n=N}^{\infty} \frac{c}{n^{1+p/2}}\left(\frac{\sqrt{2n}\pi}{2}\right)^p \\ &= c\left(\frac{\sqrt{2}\pi}{2}\right)^p \sum_{n=N}^{\infty} \frac{1}{n} = \infty,\end{aligned}$$

which contradicts the assumed bound.

□

A.4 Proof of Proposition 2

Proof. We verify the outer and inner limit conditions.

Outer limit condition. Assume that

$$(x_m, A_m(x_m)) \to (x, y) \qquad \text{in } H \times H.$$

Since $x_m \to x$, the set

$$K := \{x\} \cup \{x_m : m \in \mathbb{N}\}$$

is compact in H. By the local uniform convergence on K (note that $D(A) = D(A_m)$),

$$\sup_{z \in K \cap D(A) \cap D(A_m)} \|A_m(z) - A(z)\|_H \to 0.$$

In particular,

$$\|A_m(x_m) - A(x_m)\|_H \to 0.$$

Since also $A_m(x_m) \to y$, it follows that

$$A(x_m) \to y.$$

Because $x_m \to x$ and A is closed, we conclude that $x \in D(A)$ and $y = A(x)$. Thus the outer limit condition holds.

Inner limit condition. Let $x \in D(A)$ be arbitrary. Set

$$x_m := x \qquad (m \in \mathbb{N}).$$

Then $x_m \to x$. Since $\{x\} \subset D(A)(= D(A_m))$ is compact in H, the assumed local uniform convergence yields

$$\|A_m(x) - A(x)\|_H \to 0.$$

Hence

$$A_m(x_m) = A_m(x) \to A(x),$$

and therefore

$$(x_m, A_m(x_m)) \to (x, A(x)).$$

So the inner limit condition holds.

Combining the two parts, we obtain

$$A_m \xrightarrow{g} A.$$

□

A.5 Proof of Example 4

Proof. First, A is monotone. Indeed, if $x < y$, then every element of $A(x)$ is less than or equal to every element of $A(y)$. Hence, for $(x, u), (y, v) \in \mathrm{gph}(A)$,

$$(u - v)(x - y) \geq 0.$$

Thus A is monotone.

The graph of A is

$$\mathrm{gph}(A) = (-\infty, 0) \times \{0\} \cup \{(0, 0), (0, 1)\} \cup (0, \infty) \times \{1\},$$

which is closed in $\mathbb{R}^2$. Hence A is closed.

It is not maximally monotone, since it is properly contained in the monotone operator

$$\widetilde{A}(x) := \begin{cases} \{0\}, & x < 0, \\ [0, 1], & x = 0, \\ \{1\}, & x > 0. \end{cases}$$

Now suppose, for contradiction, that there exists a sequence of continuous maps $A_m : \mathbb{R} \to \mathbb{R}$ such that

$$A_m \xrightarrow{g} A.$$

Since $(0, 0), (0, 1) \in \mathrm{gph}(A)$, the inner condition of graph convergence gives sequences $x_m, y_m \in \mathbb{R}$ such that

$$x_m \to 0, \qquad A_m(x_m) \to 0,$$

and

$$y_m \to 0, \qquad A_m(y_m) \to 1.$$

For sufficiently large m, we have

$$A_m(x_m) < \frac{1}{2} < A_m(y_m).$$

Since A_m is continuous, the intermediate value theorem implies that there exists z_m between x_m and y_m such that

$$A_m(z_m) = \frac{1}{2}.$$

Since $x_m, y_m \to 0$, we also have $z_m \to 0$. Therefore

$$(z_m, A_m(z_m)) \to \left(0, \frac{1}{2}\right).$$

By the outer condition of graph convergence, this implies

$$\left(0, \frac{1}{2}\right) \in \mathrm{gph}(A).$$

But $A(0) = \{0, 1\}$, so $\frac{1}{2} \notin A(0)$, a contradiction. Thus no such continuous sequence exists. □

A.6 Proof of Proposition 3

Proof. **(⇒) Graph convergence implies local distance convergence.**

Assume that $A_m \xrightarrow{g} A$. Suppose, to the contrary, that there exist a compact set $K \subset H \times H$, a number $\varepsilon > 0$, and a subsequence, still denoted by m, such that

$$d_{\mathrm{gph},K}(A_m, A) \not\to 0.$$

Since

$$d_{\mathrm{gph},K}(A_m, A) \geq \varepsilon$$

along a subsequence, after passing to a further subsequence if necessary, one of the following two alternatives holds for all m:

$$(A) \quad h_K(\mathrm{gph}(A), \mathrm{gph}(A_m)) \geq \varepsilon,$$

or

$$(B) \quad h_K(\mathrm{gph}(A_m), \mathrm{gph}(A)) \geq \varepsilon.$$

Case (A): contradiction with the inner condition.

Assume (A). Then for each m there exists $(u_m, y_m) \in \mathrm{gph}(A) \cap K$ such that

$$\inf_{(v,z)\in\mathrm{gph}(A_m)} \|(u_m, y_m) - (v, z)\|_{H\times H} \geq \varepsilon/2. \tag{3}$$

Since K is compact, passing to a further subsequence, we may assume that

$$(u_m, y_m) \to (u, y).$$

Since $\mathrm{gph}(A)$ is closed, we have $(u, y) \in \mathrm{gph}(A)$. By the inner condition of graph convergence, there exists $(v_m, z_m) \in \mathrm{gph}(A_m)$ such that

$$(v_m, z_m) \to (u, y).$$

Hence

$$\begin{aligned}\inf_{(v,z)\in\mathrm{gph}(A_m)} \|(u_m, y_m) - (v, z)\|_{H\times H} &\leq \|(u_m, y_m) - (v_m, z_m)\|_{H\times H} \\ &\leq \|(u_m, y_m) - (u, y)\|_{H\times H} + \|(v_m, z_m) - (u, y)\|_{H\times H} \to 0,\end{aligned}$$

which contradicts (3).

Case (B): contradiction with the outer condition.

Assume (B). Then for each m there exists $(v_m, z_m) \in \mathrm{gph}(A_m) \cap K$ such that

$$\inf_{(u,y)\in \mathrm{gph}(A)} \|(u,y) - (v_m, z_m)\|_{H\times H} \geq \varepsilon/2.$$

Since $(v_m, z_m) \in K$ and K is compact, there exists a subsequence (not relabeled) such that

$$(v_m, z_m) \to (v, z).$$

By the outer limit condition, $(v, z) \in \mathrm{gph}(A)$. On the other hand, by continuity of the distance,

$$\inf_{(u,y)\in \mathrm{gph}(A)} \|(u,y) - (v,z)\|_{H\times H} = \lim_{m\to\infty} \inf_{(u,y)\in \mathrm{gph}(A)} \|(u,y) - (v_m, z_m)\|_{H\times H} \geq \varepsilon/2,$$

which contradicts $(v, z) \in \mathrm{gph}(A)$.

Thus $d_{\mathrm{gph},K}(A_m, A) \to 0$.

($\Leftarrow$) Local distance convergence implies graph convergence.

Assume that for every compact $K \subset H \times H$

$$\max\{h_K(\mathrm{gph}(A), \mathrm{gph}(A_m)), h_K(\mathrm{gph}(A_m), \mathrm{gph}(A))\} \to 0.$$

(Outer limit condition)

Assume $(x_m, y_m) \in \mathrm{gph}(A_m)$ and

$$(x_m, y_m) \to (x, y).$$

Define a compact set

$$K := \{(x, y)\} \cup \{(x_m, y_m)\}_{m\in\mathbb{N}}.$$

Hence

$$\inf_{(u,z)\in \mathrm{gph}(A)} \|(u,z) - (x_m, y_m)\|_{H\times H} \leq h_K(\mathrm{gph}(A_m), \mathrm{gph}(A)) \to 0.$$

Passing to the limit yields $(x, y) \in \mathrm{gph}(A)$.

(Inner limit condition)

Let $(x, y) \in \mathrm{gph}(A)$. Define a compact

$$K = \{(x, y)\}$$

Then

$$h_K(\mathrm{gph}(A), \mathrm{gph}(A_m)) \to 0,$$

so

$$\inf_{(v,z)\in \mathrm{gph}(A_m)} \|(x,y) - (v,z)\|_{H\times H} \to 0.$$

Thus we can choose $(x_m, y_m) \in \mathrm{gph}(A_m)$ such that

$$(x_m, y_m) \to (x, y).$$

Therefore $A_m \xrightarrow{g} A$.

□

A.7 Proof of Proposition 6

Proof. By the same computation of [45, Lem. 12.14], the Yosida approximation

$$A_\lambda = (\lambda I + A^{-1})^{-1}.$$

Since A^{-1} is maximally monotone, it is closed; see [8, Prop. 20.22], then by Lemma 1

$$\lambda I + A^{-1} \xrightarrow{g} A^{-1}$$

By Lemma 2 that is the continuity of the inverse operation with respect to graph convergence,

$$A_\lambda = (\lambda I + A^{-1})^{-1} \xrightarrow{g} (A^{-1})^{-1} = A.$$

□

Lemma 1. *Let $A : D(A) \subset H \to 2^H$ be closed. Then,*

$$\lambda_m I + A \xrightarrow{g} A.$$

as $\lambda_m \to 0$.

Proof. **Step 1: Outer limit condition.** Assume that $x_m \to x$, $y_m \to y$ and $y_m \in (\lambda_m I + A)(x_m)$ for all m. By the definition of $\lambda_m I + A$, we can choose $a_m \in A(x_m)$ such that

$$y_m = \lambda_m x_m + a_m.$$

Hence

$$a_m = y_m - \lambda_m x_m \to y - 0 = y.$$

Therefore $(x_m, a_m) \to (x, y)$ with $(x_m, a_m) \in \mathrm{gph}(A)$ for all m. Since A has a closed graph in $H \times H$, and we conclude that $(x, y) \in \mathrm{gph}(A)$, i.e. $y \in A(x)$.

Step 2: Inner limit condition. Let $(x, y) \in \mathrm{gph}(A)$ be arbitrary. For each m, set

$$x_m := x, \qquad y_m := \lambda_m x + y.$$

Then clearly $x_m \to x$ and $y_m \to y$ as $m \to \infty$. Moreover, since $y \in A(x)$, we have $y_m = \lambda_m x + y \in \lambda_m x + A(x) = (\lambda_m I + A)(x_m)$. Thus $(x_m, y_m) \in \mathrm{gph}(\lambda_m I + A)$ for all m and $(x_m, y_m) \to (x, y)$.

□

Lemma 2. *Let $A_m : D(A_m) \subset H \to 2^H$ and $A \subset D(A) : H \to 2^H$ such that $A_m \xrightarrow{g} A$. Then,*

$$A_m^{-1} \xrightarrow{g} A^{-1}.$$

Proof. **Step 1: Outer limit condition for A_m^{-1}.** Assume that $(y_m, x_m) \in \mathrm{gph}(A_m^{-1})$ for all m and

$$(y_m, x_m) \to (y, x) \quad \text{in } H \times H.$$

By definition of the inverse graph, $(x_m, y_m) \in \mathrm{gph}(A_m)$ for all m and

$$(x_m, y_m) \to (x, y).$$

Applying the outer condition in $A_m \xrightarrow{g} A$ yields $(x, y) \in \mathrm{gph}(A)$, i.e. $y \in A(x)$. Equivalently, $x \in A^{-1}(y)$, namely $(y, x) \in \mathrm{gph}(A^{-1})$.

Step 2: Inner limit condition for A_m^{-1}. Let $(y, x) \in \mathrm{gph}(A^{-1})$ be arbitrary. Then $(x, y) \in \mathrm{gph}(A)$. By the inner condition in $A_m \xrightarrow{g} A$, there exist $(x_m, y_m) \in \mathrm{gph}(A_m)$ such that

$$(x_m, y_m) \to (x, y).$$

Then $(y_m, x_m) \in \mathrm{gph}(A_m^{-1})$ for all m, and

$$(y_m, x_m) \to (y, x).$$

This proves the inner limit condition.

□

A.8 Proof of Theorem 1

Proof. Let $(\mu_n)_{n\in\mathbb{N}} \subset (0,\infty)$ be such that $\mu_n \to 0$. Let $K \subset H \times H$ be compact. Set

$$K_0 := \pi_1(K),$$

where $\pi_1 : H \times H \to H$ is the first projection. Set

$$K_1 := \overline{\{x + \mu_n y : (x,y) \in \mathrm{gph}(A) \cap K,\ n \in \mathbb{N}\}}, \qquad K_* := K_0 \cup K_1.$$

Then K_* is compact.

For each fixed n, since $A_{\mu_n} : H \to H$ is continuous, Proposition 7 yields

$$\inf_{\varphi\in\mathcal{F}} \sup_{u\in K_*} \|D_m \circ \varphi \circ E_m(u) - A_{\mu_n}(u)\|_H \to 0 \qquad (m \to \infty).$$

Hence, by a diagonal argument, there exists a map $n(m) \to \infty$ and $\varphi_m \in \mathcal{F}$ such that

$$\sup_{u\in K_*} \|D_m \circ \varphi_m \circ E_m(u) - A_{\mu_{n(m)}}(u)\|_H \to 0.$$

Set $\lambda_m := \mu_{n(m)}$. Then $\lambda_m \to 0$, and

$$\sup_{u\in K_*} \|A_m(u) - A_{\lambda_m}(u)\|_H \to 0. \tag{4}$$

We prove that

$$d_{\mathrm{gph},K}(A_m, A) \to 0.$$

First, we show

$$h_K(\mathrm{gph}(A_m), \mathrm{gph}(A)) \to 0.$$

Suppose not. Then there exist $\varepsilon > 0$, a subsequence, still denoted by m, and points $x_m \in H$ such that

$$(x_m, A_m(x_m)) \in K$$

and

$$\mathrm{dist}\big((x_m, A_m(x_m)), \mathrm{gph}(A)\big) \geq \varepsilon.$$

Since K is compact, up to a further subsequence,

$$(x_m, A_m(x_m)) \to (x,y) \in K.$$

In particular, $x_m \in K_0 \subset K_*$. Hence, by (4),

$$\|A_m(x_m) - A_{\lambda_m}(x_m)\|_H \to 0.$$

Therefore

$$(x_m, A_{\lambda_m}(x_m)) \to (x,y).$$

Since $A_{\lambda_m} \xrightarrow{g} A$ as $\lambda_m \to 0$, the outer limit condition gives

$$(x,y) \in \mathrm{gph}(A).$$

This contradicts

$$\mathrm{dist}\big((x_m, A_m(x_m)), \mathrm{gph}(A)\big) \geq \varepsilon.$$

Hence

$$h_K(\mathrm{gph}(A_m), \mathrm{gph}(A)) \to 0.$$

Next, we show

$$h_K(\mathrm{gph}(A), \mathrm{gph}(A_m)) \to 0.$$

Let $(x,y) \in \mathrm{gph}(A) \cap K$. Define

$$z_m := x + \lambda_m y.$$

Since $y \in Ax$, we have

$$z_m = x + \lambda_m y \in x + \lambda_m Ax,$$

and hence

$$J_{\lambda_m} z_m = x.$$

Therefore

$$A_{\lambda_m}(z_m) = \frac{z_m - J_{\lambda_m} z_m}{\lambda_m} = \frac{x + \lambda_m y - x}{\lambda_m} = y.$$

Moreover, by definition of K_1, we have $z_m \in K_1 \subset K_*$. Hence, using (4),

$$\|A_m(z_m) - y\|_H = \|A_m(z_m) - A_{\lambda_m}(z_m)\|_H \to 0$$

uniformly for $(x, y) \in \mathrm{gph}(A) \cap K$. Also,

$$\|z_m - x\|_H = \lambda_m \|y\|_H \to 0$$

uniformly for $(x, y) \in K$, since K is compact and therefore bounded. Thus

$$\mathrm{dist}\big((x, y), \mathrm{gph}(A_m)\big) \le \|(x, y) - (z_m, A_m(z_m))\|_{H \times H} \to 0$$

uniformly for $(x, y) \in \mathrm{gph}(A) \cap K$. Hence

$$h_K(\mathrm{gph}(A), \mathrm{gph}(A_m)) \to 0.$$

Combining the two estimates gives

$$d_{\mathrm{gph},K}(A_m, A) \to 0.$$

This completes the proof. □

A.9 Proof of Corollary 1

Proof. Let

$$J_\lambda := (I + \lambda A)^{-1}$$

be the resolvent of A, and define

$$R_\lambda := 2J_\lambda - I.$$

Since A is maximally monotone, J_λ is firmly nonexpansive. Hence R_λ is nonexpansive, namely,

$$\|R_\lambda(x) - R_\lambda(y)\|_H \le \|x - y\|_H \qquad (x, y \in H).$$

Moreover,

$$A_\lambda := \frac{I - J_\lambda}{\lambda} = \frac{I - R_\lambda}{2\lambda}.$$

Fix a compact set $K \subset H \times H$. Set

$$K_0 := \pi_1(K),$$

where $\pi_1 : H \times H \to H$ denotes the first projection, and define

$$K_1 := \{x + \lambda y : (x, y) \in \mathrm{gph}(A) \cap K,\ \lambda \in [0, 1]\}.$$

Since $\mathrm{gph}(A)$ is closed, $\mathrm{gph}(A) \cap K$ is compact, and hence K_1 is compact. Put

$$K_* := K_0 \cup K_1.$$

Then K_* is compact.

We next choose $\lambda_m \to 0$ together with the approximating maps. For each $j \in \mathbb{N}$, set

$$\mu_j := \frac{1}{j}.$$

Since $R_{\mu_j} : H \to H$ is 1-Lipschitz, Proposition 7 yields a sequence $(\psi_m^{(j)})_{m \in \mathbb{N}} \subset \mathcal{F}^{\mathrm{Lip}}$ such that

$$\sup_{u \in K_*} \|D_m \circ \psi_m^{(j)} \circ E_m(u) - R_{\mu_j}(u)\|_H \to 0 \qquad (m \to \infty).$$

Therefore, for each $j \in \mathbb{N}$, there exists $M_j \in \mathbb{N}$ such that

$$\sup_{u \in K_*} \|D_m \circ \psi_m^{(j)} \circ E_m(u) - R_{\mu_j}(u)\|_H \le \frac{1}{j^2} \qquad (m \ge M_j).$$

By increasing M_j if necessary, we may assume that

$$M_1 < M_2 < \cdots \qquad \text{and} \qquad M_j \to \infty.$$

For $m \geq M_1$, define

$$j(m) := \max\{j \in \mathbb{N} : M_j \leq m\},$$

and set

$$\lambda_m := \mu_{j(m)} = \frac{1}{j(m)}, \qquad \psi_m := \psi_m^{(j(m))}.$$

For the finitely many indices $m < M_1$, choose $\lambda_m > 0$ and $\psi_m \in \mathcal{F}^{\mathrm{Lip}}$ arbitrarily. Since $j(m) \to \infty$, we have

$$\lambda_m \to 0.$$

Moreover, for $m \geq M_1$,

$$\sup_{u \in K_*} \|D_m \circ \psi_m \circ E_m(u) - R_{\lambda_m}(u)\|_H \leq \frac{1}{j(m)^2} = \lambda_m^2 = o(\lambda_m).$$

Thus

$$\sup_{u \in K_*} \|D_m \circ \psi_m \circ E_m(u) - R_{\lambda_m}(u)\|_H = o(\lambda_m). \tag{5}$$

Set

$$S_m := D_m \circ \psi_m \circ E_m, \qquad B_m := \frac{I - S_m}{2\lambda_m}.$$

Then

$$S_m : H \to H$$

is 1-Lipschitz. We see that

$$\begin{aligned}
\sup_{u \in K_*} \|B_m(u) - A_{\lambda_m}(u)\|_H &= \sup_{u \in K_*} \left\| \frac{I - S_m}{2\lambda_m}(u) - \frac{I - R_{\lambda_m}}{2\lambda_m}(u) \right\|_H \\
&\leq \frac{1}{2\lambda_m} \sup_{u \in K_*} \|S_m(u) - R_{\lambda_m}(u)\|_H \\
&\to 0.
\end{aligned}$$

Hence

$$\sup_{u \in K_*} \|B_m(u) - A_{\lambda_m}(u)\|_H \to 0. \tag{6}$$

We prove that

$$d_{\mathrm{gph},K}(B_m, A) \to 0.$$

First, we show

$$h_K(\mathrm{gph}(B_m), \mathrm{gph}(A)) \to 0.$$

Suppose not. Then there exist $\varepsilon > 0$, a subsequence still denoted by m, and points $x_m \in H$ such that

$$(x_m, B_m(x_m)) \in K$$

and

$$\mathrm{dist}\big((x_m, B_m(x_m)), \mathrm{gph}(A)\big) \geq \varepsilon.$$

Since K is compact, after passing to a further subsequence,

$$(x_m, B_m(x_m)) \to (x, y) \in K.$$

Since $x_m \in K_0 \subset K_*$, (6) implies

$$\|B_m(x_m) - A_{\lambda_m}(x_m)\|_H \to 0.$$

Hence

$$(x_m, A_{\lambda_m}(x_m)) \to (x, y).$$

Since $A_{\lambda_m} \xrightarrow{g} A$ as $\lambda_m \to 0$, the outer limit condition yields

$$(x, y) \in \mathrm{gph}(A),$$

which contradicts the assumption. Therefore

$$h_K(\mathrm{gph}(B_m), \mathrm{gph}(A)) \to 0.$$

Next, we show

$$h_K(\mathrm{gph}(A), \mathrm{gph}(B_m)) \to 0.$$

Let

$$(x, y) \in \mathrm{gph}(A) \cap K.$$

Define

$$z_m := x + \lambda_m y.$$

Since $y \in Ax$,

$$z_m \in x + \lambda_m Ax,$$

and therefore

$$J_{\lambda_m} z_m = x.$$

Hence

$$R_{\lambda_m}(z_m) = 2x - z_m = x - \lambda_m y.$$

Consequently,

$$A_{\lambda_m}(z_m) = \frac{z_m - J_{\lambda_m} z_m}{\lambda_m} = y.$$

For all sufficiently large m, we have $\lambda_m \in [0, 1]$, and hence $z_m \in K_1 \subset K_*$. Thus (6) gives

$$\|B_m(z_m) - y\|_H = \|B_m(z_m) - A_{\lambda_m}(z_m)\|_H \to 0$$

uniformly for $(x, y) \in \mathrm{gph}(A) \cap K$. Moreover,

$$\|z_m - x\|_H = \lambda_m \|y\|_H \to 0$$

uniformly for $(x, y) \in K$, since K is bounded.

Therefore

$$\mathrm{dist}\big((x, y), \mathrm{gph}(B_m)\big) \le \|(x, y) - (z_m, B_m(z_m))\|_{H \times H} \to 0$$

uniformly for

$$(x, y) \in \mathrm{gph}(A) \cap K.$$

Hence

$$h_K(\mathrm{gph}(A), \mathrm{gph}(B_m)) \to 0.$$

Combining the above two estimates,

$$d_{\mathrm{gph},K}(B_m, A) \to 0.$$

Finally, we show that B_m is maximally monotone. Since S_m is 1-Lipschitz, for any $x, y \in H$,

$$\begin{aligned}
\langle B_m x - B_m y, x - y \rangle &= \frac{1}{2\lambda_m} \Big(\|x - y\|_H^2 - \langle S_m x - S_m y, x - y \rangle \Big) \\
&\ge \frac{1}{2\lambda_m} \Big(\|x - y\|_H^2 - \|S_m x - S_m y\|_H \|x - y\|_H \Big) \\
&\ge 0.
\end{aligned}$$

Thus B_m is monotone. Since $B_m : H \to H$ is continuous and everywhere defined, it is maximally monotone. □

B Examples of maximally monotone operators

We give some examples of maximally monotone operators.

(a) Subdifferentials of convex functionals $H \to \mathbb{R} \cup \{+\infty\}$ such as

- the p-Laplacian, $1 < p < +\infty$, $A : W^{1,p}(\Omega) \subset L^2(\Omega) \to L^2(\Omega)$ for a bounded open set $\Omega \subset \mathbb{R}^d$, defined as $A(u) = -\mathrm{div}(|\nabla u|^{p-2}\nabla u)$, which is a subgradient of the p-Dirichlet energy functional $J \colon L^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ given by

$$J(u) = \begin{cases} \frac{1}{p}\int_\Omega \|\nabla u\|^p \, dx, & u \in W^{1,p}(\Omega), \\ +\infty, & u \in L^2(\Omega) \setminus W^{1,p}(\Omega). \end{cases}$$

- the 1-Laplacian $A : \mathrm{BV}(\Omega)(\Omega) \subset L^2(\Omega) \to L^2(\Omega)$, where $\mathrm{BV}(\Omega)$ is the space of functions of bounded variation, formally written as $A(u) = -\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right)$ but rigorously defined as the subgradient of minimal norm of the Total Variation functional $J \colon L^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$

$$J(u) = \begin{cases} \sup\left\{\int_\Omega u \operatorname{div} \phi \, dx \, : \, \phi \in C_c^\infty(\Omega; \mathbb{R}^n), \, \sup_\Omega |\phi| \le 1\right\}, & u \in \mathrm{BV}(\Omega), \\ +\infty, & u \in L^2(\Omega) \setminus \mathrm{BV}(\Omega). \end{cases}$$

- the ∞-Laplacian [5, 30], which can be formally written as $\Delta_\infty u := \langle \nabla u, D^2 u \nabla u \rangle$ for sufficiently smooth functions but is rigorously defined in the viscosity sense; it is also a subgradient of the Lipschitz constant functional defined on the Banach space $C(\Omega)$

$$J(u) = \begin{cases} \|\nabla u\|_{L^\infty}, & u \in W^{1,\infty}(\Omega), \\ +\infty, & u \in C(\Omega) \setminus W^{1,\infty}(\Omega) \end{cases}$$

  and, similarly to the p-Laplacian, has a divergence form $A(u) = -\operatorname{div}(\tau \nabla_\tau u)$ where τ is a certain measure that depends on u and ∇_τ is the tangential gradient [19].
- the operator $u \mapsto \Delta(u^{[m]})$ where $m > 0$ is a constant and $u^{[m]} := \mathrm{sign}(u)|u|^m$, which arises in the porous medium ($m > 1$) and the fast diffusion ($m < 1$) equations; it is a subgradient of the following functional defined on the negative-order Sobolev space $H^{-1}(\Omega)$ (the dual of the Sobolev space $H_0^1(\Omega)$ with homogeneous Dirichlet boundary conditions)

$$J(u) := \begin{cases} \frac{1}{m+1}\int_\Omega |u|^{m+1} \, dx, & \text{if } u \in L^{m+1}(\Omega) \cap H^{-1}(\Omega), \\ \infty, & u \in H^{-1}(\Omega) \setminus L^{m+1}(\Omega). \end{cases}$$

  More details can be found, for example, in [17, 18, 20].

(b) The differentiation operator (the inverse Volterra operator [7]) on periodic functions $A : H^1_{\mathrm{per}}(0,1) \subset L^2_{\mathrm{per}}(0,1) \to L^2_{\mathrm{per}}(0,1)$, $A(u)(x) = \frac{d}{dx}u(x)$ (which is not symmetric and therefore cannot be the subdifferential of a convex function).

(c) The inverse A^{-1} of a maximally monotone operator A is maximally monotone; see [8, Prop. 20.22]. Thus, the inverses of the above operators are also maximally monotone.

C Quantitative result

Using quantitative approximation results for neural networks applied to the Yosida approximation, we obtain the following.

Corollary 2 (Quantitative local graph estimate)**.** *Assume that H has the EDAP with encoder and decoder mappings $E_m : H \to \mathbb{R}^{w_m}$ and $D_m : \mathbb{R}^{w_m} \to H$. Let $A : D(A) \subset H \to 2^H$ be maximally monotone and assume that $0 \in D(A)$. Let $K = K_1 \times K_2 \subset H \times H$ be compact with $0 \in K_1$. For $\lambda \in (0,1)$ set*

$$S_\lambda(K) := K_1 \cup \{x + \lambda y : (x,y) \in \mathrm{gph}(A) \cap K\}.$$

Then, for every $\lambda \in (0,1)$ and $m \in \mathbb{N}$, there exists a ReLU network $\varphi_{\lambda,m,N} : \mathbb{R}^{w_m} \to \mathbb{R}^{w_m}$ with size parameter $N \in \mathbb{N}$ such that, with

$$B_{\lambda,m,N} := D_m \circ \varphi_{\lambda,m,N} \circ E_m,$$

one has

$$d_{\mathrm{gph},K}(B_{\lambda,m,N},A) \le \lambda R_K + (1+\lambda)\Bigg[\frac{1}{\lambda}C_{S_\lambda(K)}(w_m) + C_{A_\lambda\circ D_m\circ E_m(S_\lambda(K))}(w_m) + C_{\mathrm{Yar}}(w_m)\left(\|A^0(0)\|_H + \frac{1}{\lambda}\big(1+\mathrm{diam}(S_\lambda(K))\big)\right)\mathrm{diam}(S_\lambda(K))N^{-1/w_m}\Bigg],$$

where

$$R_K := \sup_{y\in K_2}\|y\|_H, \qquad C_B(w_m) := \sup_{u\in B}\|u - D_m\circ E_m(u)\|_H.$$

Corollary 2 highlights the trade-off between λ, m, and N. The first term vanishes as $\lambda \to 0$, whereas the second and third terms and the fourth term may increase. On the other hand, for fixed $\lambda > 0$, the second and third terms vanish as $m \to \infty$. Moreover, for fixed $\lambda > 0$ and $m \in \mathbb{N}$, the fourth term vanishes as $N \to \infty$.

Remark 3. *We give several remarks on the upper bound estimates obtained in the above.*

(a) Although we obtain an upper bound on the local graph distance between the target operator and the encoder–decoder architecture, some terms are not fully explicit. While the constant $C_{\mathrm{Yar}}(w_m)$ arising from the Yarotsky-type approximation theorem is generally not explicit, the encoder–decoder errors $C_B(w_m)$ may admit explicit bounds under additional structural assumptions on the encoder and decoder mappings E_m, D_m.

(b) We emphasize that although N does not depend on the compact, the distance $d_{\mathrm{gph},K}$, in which the rate is obtained, does depend on K (see Definition 5). Therefore, despite its appearance, this error estimate is not uniform over compact sets.

Proof. Set

$$B_{\lambda,m,N} := D_m \circ \varphi_{\lambda,m,N} \circ E_m.$$

Define

$$\eta_{\lambda,m,N} := \sup_{u\in S_\lambda(K)}\|B_{\lambda,m,N}(u) - A_\lambda(u)\|_H.$$

We claim that

$$d_{\mathrm{gph},K}(B_{\lambda,m,N},A) \le \lambda R_K + (1+\lambda)\eta_{\lambda,m,N}.$$

Indeed, let $(u, B_{\lambda,m,N}(u)) \in \mathrm{gph}(B_{\lambda,m,N}) \cap K$. Set

$$q := A_\lambda(u), \qquad p := J_\lambda u.$$

Then $q \in A(p)$ and $u = p + \lambda q$. Hence $(p,q) \in \mathrm{gph}(A)$. Moreover, since $B_{\lambda,m,N}(u) \in K_2$,

$$\|q\|_H \le \|B_{\lambda,m,N}(u)\|_H + \|q - B_{\lambda,m,N}(u)\|_H \le R_K + \eta_{\lambda,m,N}.$$

Therefore

$$\begin{aligned}\mathrm{dist}\big((u, B_{\lambda,m,N}(u)), \mathrm{gph}(A)\big) &\le \|(u, B_{\lambda,m,N}(u)) - (p,q)\|_{H\times H}\\ &\le \lambda\|q\|_H + \|B_{\lambda,m,N}(u) - q\|_H\\ &\le \lambda R_K + (1+\lambda)\eta_{\lambda,m,N}.\end{aligned}$$

Taking the supremum over $\mathrm{gph}(B_{\lambda,m,N}) \cap K$ gives

$$h_K(\mathrm{gph}(B_{\lambda,m,N}), \mathrm{gph}(A)) \le \lambda R_K + (1+\lambda)\eta_{\lambda,m,N}.$$

Conversely, let $(x,y) \in \mathrm{gph}(A) \cap K$ and set

$$z := x + \lambda y.$$

Then $z \in S_\lambda(K)$ and, by the resolvent identity,

$$J_\lambda z = x, \qquad A_\lambda(z) = y.$$

Hence

$$(z, B_{\lambda,m,N}(z)) \in \mathrm{gph}(B_{\lambda,m,N}),$$

and therefore

$$\begin{aligned}\operatorname{dist}\big((x,y),\operatorname{gph}(B_{\lambda,m,N})\big) &\le \|(x,y)-(z,B_{\lambda,m,N}(z))\|_{H\times H}\\ &\le \lambda\|y\|_H + \|y-B_{\lambda,m,N}(z)\|_H\\ &\le \lambda R_K + \eta_{\lambda,m,N}.\end{aligned}$$

Taking the supremum over $\operatorname{gph}(A)\cap K$ yields

$$h_K(\operatorname{gph}(A),\operatorname{gph}(B_{\lambda,m,N})) \le \lambda R_K + \eta_{\lambda,m,N}.$$

Combining the two one-sided estimates gives

$$d_{\operatorname{gph},K}(B_{\lambda,m,N},A) \le \lambda R_K + (1+\lambda)\eta_{\lambda,m,N}.$$

It remains to bound $\eta_{\lambda,m,N}$. For $u\in S_\lambda(K)$, we decompose

$$\begin{aligned}\|A_\lambda(u)-B_{\lambda,m,N}(u)\|_H \le{}& \|A_\lambda(u)-A_\lambda(D_mE_mu)\|_H\\ &+\|A_\lambda(D_mE_mu)-D_mE_mA_\lambda(D_mE_mu)\|_H\\ &+\|D_mE_mA_\lambda(D_mE_mu)-D_m\varphi_{\lambda,m,N}E_mu\|_H.\end{aligned}$$

Since A_λ is $1/\lambda$-Lipschitz, the first term is bounded by

$$\frac{1}{\lambda}C_{S_\lambda(K)}(w_m).$$

The second term is bounded by

$$C_{A_\lambda\circ D_m\circ E_m(S_\lambda(K))}(w_m).$$

For the third term, note that $E_m\circ A_\lambda\circ D_m$ is Lipschitz on $E_m(S_\lambda(K))$, and hence belongs to $W^{1,\infty}(E_m(S_\lambda(K));\mathbb{R}^{w_m})$. By the Yarotsky-type ReLU approximation theorem [48, Thm. 1], there exists a ReLU network $\varphi_{\lambda,m,N}:\mathbb{R}^{w_m}\to\mathbb{R}^{w_m}$ such that

$$\begin{aligned}\sup_{z\in E_m(S_\lambda(K))}\|E_m\circ A_\lambda\circ D_m(z)-\varphi_{\lambda,m,N}(z)\| \le{}& C_{\mathrm{Yar}}(w_m)\|E_m\circ A_\lambda\circ D_m\|_{W^{1,\infty}(E_m(S_\lambda(K));\mathbb{R}^{w_m})}\\ &\times \operatorname{diam}(E_m(S_\lambda(K)))\,N^{-1/w_m}.\end{aligned}$$

Since $\operatorname{diam}(E_m(S_\lambda(K)))\le\operatorname{diam}(S_\lambda(K))$, we estimate (by $0\in K_1$)

$$\begin{aligned}\|E_m\circ A_\lambda\circ D_m\|_{W^{1,\infty}} &\le \|E_m\circ A_\lambda\circ D_m\|_{L^\infty}+\operatorname{Lip}(E_m\circ A_\lambda\circ D_m)\\ &\le \|A^0(0)\|_H+\frac{1}{\lambda}\operatorname{diam}(S_\lambda(K))+\frac{1}{\lambda}.\end{aligned}$$

Thus,

$$\begin{aligned}&\sup_{z\in E_m(S_\lambda(K))}\|E_m\circ A_\lambda\circ D_m(z)-\varphi_{\lambda,m,N}(z)\|\\ &\qquad\le C_{\mathrm{Yar}}(w_m)\left(\|A^0(0)\|_H+\frac{1}{\lambda}(1+\operatorname{diam}(S_\lambda(K)))\right)\operatorname{diam}(S_\lambda(K_1))\,N^{-1/w_m}.\end{aligned}$$

Since D_m is 1-Lipschitz, the same bound applies after composing with D_m. Substituting these estimates into the deterministic local graph estimate proves the result. □

D Experiments

D.1 Experiments for differential operators

We study operator learning for the following two differential operators:

$$\begin{aligned}&A_1 : u(t)\mapsto u'(t),\\ &A_2 : u(x,y)\mapsto -\operatorname{div}\big(|\nabla u(x,y)|^{p-2}\nabla u(x,y)\big),\end{aligned}$$

defined on $[0,1]$ and $[0,1]^2$, respectively. We generate input data as

$$u(t) = \sum_{j=1}^{K} \frac{1}{n_j^{\beta}} \Big(a_j \sin(2\pi n_j t) + b_j \cos(2\pi n_j t) \Big), \tag{7}$$

$$u(x,y) = \sum_{j=1}^{K} \frac{1}{(\sqrt{k_j^2 + \ell_j^2})^{\beta}} \Big[a_j \sin(2\pi k_j x)\sin(2\pi \ell_j y) + b_j \sin(2\pi k_j x)\cos(2\pi \ell_j y) + c_j \cos(2\pi k_j x)\sin(2\pi \ell_j y) + d_j \cos(2\pi k_j x)\cos(2\pi \ell_j y) \Big]. \tag{8}$$

Here,

$$n_j \sim \mathrm{Uniform}\{n_{\min}^{\mathrm{train}}, \dots, n_{\max}^{\mathrm{train}}\}$$

for the training data, and

$$n_j \sim \mathrm{Uniform}\{n_{\min}^{\mathrm{test}}, \dots, n_{\max}^{\mathrm{test}}\}$$

for the test data. Moreover,

$$a_j, b_j, c_j, d_j \sim \mathcal{N}(0,1)$$

are independent random variables. The training and test datasets are constructed as

$$\{(u_j, A_k(u_j))\}_{j=1}^{N^{\mathrm{train}}} \quad \text{and} \quad \{(u_j, A_k(u_j))\}_{j=1}^{N^{\mathrm{test}}},$$

respectively, where each u_j is sampled according to (7) or (8), and where N^{train} and N^{test} denote the numbers of training and test samples.

This data generation is inspired by Examples 2 and 3, which illustrate obstructions to uniform or L^p approximation for suitable classes of inputs or input measures, thereby motivating the use of graph-distance-based losses. In particular, high-frequency inputs of the form (7) and (8) are mapped to outputs with significantly amplified amplitudes, reflecting the unbounded nature of differential operators (see Figures 1 and 2).

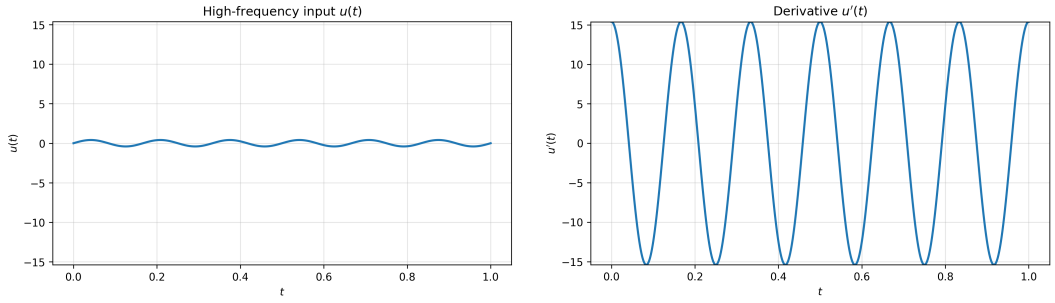

Figure 1: High-frequency input $u(t)$ (left) and its derivative $u'(t)$ (right). Here $u(t)$ is generated as in (7) with $K = 1$, $n = 6$, $a_j = 1$, $b_j = 0$, and $\beta = 0.5$.

We compare several models $\hat{A}$ based on Fourier Neural Operators (FNOs), trained with different loss functions:

$$\sum_{j=1}^{N^{\mathrm{train}}} \|A_k(u_j) - \hat{A}(u_j)\|_{L^2}^2 \qquad (\ell^2 \text{ loss}), \tag{9}$$

$$\max_{j \in [N^{\mathrm{train}}]} \|A_k(u_j) - \hat{A}(u_j)\|_{L^2} \;\approx\; \tau_\infty \log\left(\sum_{j=1}^{N^{\mathrm{train}}} \exp\left(\frac{\|A_k(u_j) - \hat{A}(u_j)\|_{L^2}}{\tau_\infty} \right) \right) \qquad (\text{soft } \ell^\infty \text{ loss}), \tag{10}$$

and a graph-distance-based loss defined via the soft approximation

$$\max\Big\{ \sup_{i \in [N^{\mathrm{train}}]} \inf_{j \in [N^{\mathrm{train}}]} d_{ij}, \sup_{j \in [N^{\mathrm{train}}]} \inf_{i \in [N^{\mathrm{train}}]} d_{ij} \Big\} \approx \max\{H, H'\} =: d_{\text{soft-gph}}(A_k, \hat{A}) \qquad (\text{soft graph loss}), \tag{11}$$

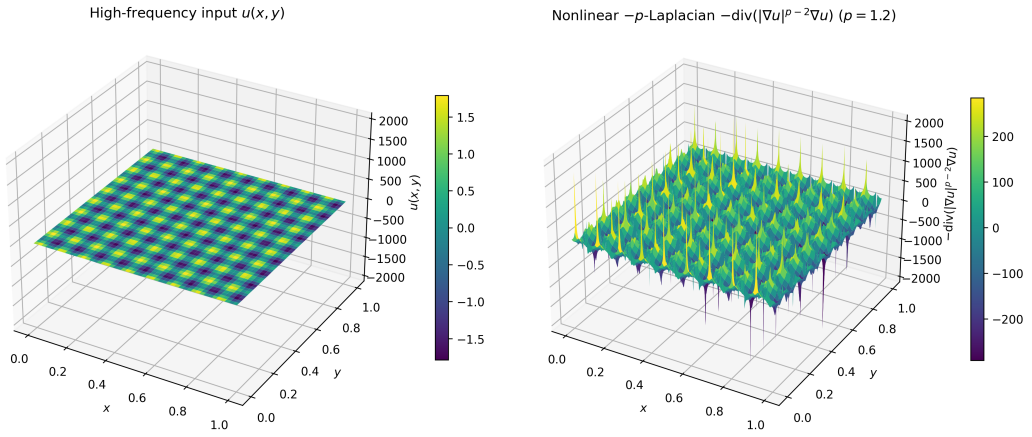

Figure 2: High-frequency input $u(x, y)$ (left) and the corresponding nonlinear p-Laplacian $-\mathrm{div}(|\nabla u|^{p-2}\nabla u)$ with $p = 1.2$ (right). Here $u(x, y)$ is generated as in (8) with $K = 1$, $n = 9$, $a_j = 1$, $b_j = 1$, and $\beta = 0$.

with weighted distance

$$d_{ij} = \sqrt{w_1 \|u_i - u_j\|_{L^2}^2 + w_2 \|A_k(u_i) - \hat{A}(u_j)\|_{L^2}^2}.$$

The soft approximations are defined by

$$H = \tau_{\mathrm{gph,out}} \log \sum_{i=1}^{N^{\mathrm{train}}} \exp\left(\frac{-\tau_{\mathrm{gph,in}} \log \sum_{j=1}^{N^{\mathrm{train}}} \exp(-d_{ij}/\tau_{\mathrm{gph,in}})}{\tau_{\mathrm{gph,out}}} \right),$$

and

$$H' = \tau_{\mathrm{gph,out}} \log \sum_{j=1}^{N^{\mathrm{train}}} \exp\left(\frac{-\tau_{\mathrm{gph,in}} \log \sum_{i=1}^{N^{\mathrm{train}}} \exp(-d_{ij}/\tau_{\mathrm{gph,in}})}{\tau_{\mathrm{gph,out}}} \right).$$

The soft graph loss is a computable finite-sample surrogate motivated by the local graph discrepancy introduced in Definition 5. The theoretical results of this paper concern the latter quantity. We do not establish a formal equivalence or consistency result between the finite-sample soft graph loss and the theoretical local graph discrepancy.

For a fair comparison, all models use the same FNO architecture, including the same number of modes, width, and depth. All models are trained using Adam.

In addition, we introduce a structured model inspired by the resolvent formulation of maximally monotone operators (Corollary 1). Instead of directly parameterizing $\hat{A}$, we define

$$\hat{S}(u) = u - \lambda \hat{R}(u), \qquad \hat{A}(u) = \frac{u - \hat{S}(u)}{\lambda}, \tag{12}$$

where $\hat{R}$ is represented by an FNO. Note that this parametrization does not restrict the expressive power of $\hat{A}$, since formally $\hat{A}(u) = \hat{R}(u)$. However, it allows us to impose structural constraints on $\hat{S}$, which, if enforced exactly in the appropriate reflected-resolvent construction, can ensure maximal monotonicity. The sampled penalty below does not itself provide such a guarantee. To promote this structure, we add a penalty encouraging $\hat{S}$ to be non-expansive:

$$\mathcal{L}_{\mathrm{nonexp}} = \sum_{(p,q)} \left[\max\left\{ 0, \frac{\|\hat{S}(u_p) - \hat{S}(u_q)\|_{L^2}}{\|u_p - u_q\|_{L^2}} - 1 \right\}^2 \right],$$
$$\text{where } u_p, u_q \sim (u_j)_{j=1}^{N^{\mathrm{train}}} \quad \text{with} \quad \#\{(u_p, u_q)\} = N^{\mathrm{nonexp}}.$$

This term penalizes violations of the non-expansive condition

$$\|\hat{S}(x) - \hat{S}(y)\|_{L^2} \le \|x - y\|_{L^2}.$$

The structure-regularized model is trained using the loss

$$\mathcal{L} = d_{\text{soft-gph}}(A, \hat{A}) + \gamma \mathcal{L}_{\text{nonexp}}.$$

In total, we train the following four models:

- FNO trained with ℓ^2 loss,
- FNO trained with soft ℓ^∞ loss,
- FNO trained with soft graph loss,
- Structured FNO trained with soft graph loss and non-expansive regularization.

We evaluate the trained models using the following metrics:

- **Test MSE**

$$\text{Test MSE} := \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \|\hat{A}(u_i) - A_k(u_i)\|_{L^2}^2. \tag{13}$$

- **Mean Rel-L2**

$$\text{Mean Rel-L2} := \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \frac{\|\hat{A}(u_i) - A_k(u_i)\|_{L^2}}{\|A_k(u_i)\|_{L^2}}. \tag{14}$$

- **Worst Rel-L2**

$$\text{Worst Rel-L2} := \max_{1 \le i \le N^{\text{test}}} \frac{\|\hat{A}(u_i) - A_k(u_i)\|_{L^2}}{\|A_k(u_i)\|_{L^2}}. \tag{15}$$

- **Test Graph**

$$\text{Test Graph} := \max\Big\{ \sup_{i \in [N^{\text{test}}]} \inf_{j \in [N^{\text{test}}]} d_{ij}, \sup_{j \in [N^{\text{test}}]} \inf_{i \in [N^{\text{test}}]} d_{ij} \Big\} \tag{16}$$

- **Mono mean-viol**
  For evaluated pairs (u_p, u_q), define

$$m_{pq} := \big\langle u_p - u_q, \hat{A}(u_p) - \hat{A}(u_q) \big\rangle_H, \tag{17}$$

  where

$$u_p, u_q \in (u_j)_{j=1}^{N^{\text{test}}}, \qquad \#\{(u_p, u_q)\} = N^{\text{mono}}.$$

  The mean monotonicity violation is defined by

$$\text{Mono mean-viol} := \frac{1}{N^{\text{mono}}} \sum_{(p,q)} \max(-m_{pq}, 0). \tag{18}$$

- **Mono worst-margin**
  The smallest signed monotonicity margin is defined by

$$\text{Mono worst-margin} := \min_{(p,q)} m_{pq}. \tag{19}$$

- **Mono frac**
  The fraction of violating pairs is defined by

$$\text{Mono frac} := \frac{1}{N^{\text{mono}}} \sum_{(p,q)} \mathbf{1}_{\{m_{pq} < -10^{-8}\}}, \tag{20}$$

  where $\mathbf{1}$ denotes the indicator function. We use the numerical tolerance 10^{-8} when identifying violating pairs.

Tables 1 and 2 summarize the choice of hyperparameters.

The results are summarized in Tables 3 and 4 where values are reported as mean $\pm$ standard deviation over four random seeds. The A_2 test set uses frequencies outside the training range; its reported test results are therefore out-of-distribution (OOD).

Table 1: Hyperparameters for the experiment on $A_1(u) = u'$.

Parameter	Value
Number of grid points for $[0, 1]$	128
Training samples N^{train}	2000
Test samples N^{test}	400
Batch size	64
Training frequency range	$n_j \in \{2, \dots, 30\}$
Test frequency range	$n_j \in \{2, \dots, 30\}$
Number of modes per sample K	$K \sim \mathrm{Uniform}\{3, \dots, 6\}$
Frequency decay exponent β	0.5
Epochs	80
Optimizer	Adam
Learning rate	2×10^{-3}
Weight decay	10^{-6}
Gradient clipping threshold	1.0
Soft ℓ^∞ temperature τ_∞	0.02
Graph inner temperature $\tau_{\mathrm{gph,in}}$	0.01
Graph outer temperature $\tau_{\mathrm{gph,out}}$	0.01
Graph input weight w_1	0.5
Graph output weight w_2	1.0
FNO modes	34
FNO width	100
FNO depth	3
Structured model parameter λ	0.01
Non-expansive penalty weight γ	0.01
Number of non-expansive pairs N^{nonexp}	512
Monotonicity evaluation pairs N^{mono}	128^2

Table 2: Hyperparameters for the experiment on $A_2(u) = -\operatorname{div}(|\nabla u|^{p-2}\nabla u)$.

Parameter	Value
Number of grid points for $[0,1]^2$	32×32
Training samples N^{train}	2000
Test samples N^{test}	400
Batch size	32
Training frequency range	$(k,l) \in \{2,\dots,6\}^2$
OOD test frequency range	$(k,l) \in \{9,\dots,16\}^2$
Number of modes per sample K	$K \sim \mathrm{Uniform}\{1,\dots,8\}$
Frequency decay exponent β	0.0
p-Laplacian exponent p	1.2
Epochs	100
Optimizer	Adam
Learning rate	3×10^{-4}
Weight decay	10^{-6}
Gradient clipping threshold	0.3
Soft ℓ^∞ temperature τ_∞	0.1
Graph inner temperature $\tau_{\mathrm{gph,in}}$	10^{-4}
Graph outer temperature $\tau_{\mathrm{gph,out}}$	10^{-4}
Graph input weight w_1	10^{-4}
Graph output weight w_2	1.0
FNO modes	16
FNO width	96
FNO depth	3
Structured model parameter λ	7×10^{-3}
Non-expansive penalty weight γ	2×10^{-5}
Number of non-expansive pairs N^{nonexp}	64
Monotonicity evaluation pairs N^{mono}	5000

Table 3: Comparison of test evaluations for $A_1(u) = u'$.

Model	Test MSE	Mean Rel-ℓ^2	Worst Rel-ℓ^2
ℓ^2	0.3919 ± 0.4013	$9.551 \times 10^{-3} \pm 6.1 \times 10^{-3}$	0.0226 ± 0.0163
ℓ^∞	1.2375 ± 0.3760	0.0209 ± 0.0030	0.0451 ± 0.0141
graph	$\mathbf{0.0420 \pm 0.0053}$	$\mathbf{3.367 \times 10^{-3} \pm 1.9 \times 10^{-4}}$	$\mathbf{0.0116 \pm 0.0029}$
graph + structure	0.1876 ± 0.0561	$8.148 \times 10^{-3} \pm 1.6 \times 10^{-3}$	0.0153 ± 0.0014

Model	Test Graph	Mono mean-viol	Mono worst-margin	Mono frac
ℓ^2	1.9502 ± 1.8868	8.7975 ± 8.0704	-110.9443 ± 74.3604	0.5847 ± 0.2193
ℓ^∞	4.1638 ± 1.9444	8.7800 ± 7.2842	-161.1156 ± 83.9334	0.4540 ± 0.3070
graph	$\mathbf{0.2784 \pm 0.1235}$	1.8198 ± 0.3501	-33.9190 ± 8.9241	0.5276 ± 0.0896
graph + structure	0.8571 ± 0.2695	$\mathbf{0.0396 \pm 0.0465}$	$\mathbf{-18.3752 \pm 20.5491}$	$\mathbf{4.852 \times 10^{-3} \pm 5.5 \times 10^{-3}}$

Table 4: Comparison of test evaluations for $A_2(u) = -\operatorname{div}(|\nabla u|^{p-2}\nabla u)$.

Model	Test MSE	Mean Rel-ℓ^2	Worst Rel-ℓ^2
ℓ^2	$1.950 \times 10^4 \pm 2.9 \times 10^3$	0.6176 ± 0.0456	0.6917 ± 0.0362
ℓ^∞	$2.015 \times 10^4 \pm 2.7 \times 10^3$	0.6279 ± 0.0420	0.7145 ± 0.0487
graph	$\mathbf{1.557 \times 10^4 \pm 2.1 \times 10^3}$	$\mathbf{0.5516 \pm 0.0373}$	$\mathbf{0.6231 \pm 0.0436}$
graph + structure	$3.882 \times 10^4 \pm 1.0 \times 10^3$	0.8771 ± 0.0111	0.9198 ± 0.0159

Model	Test Graph	Mono mean-viol	Mono worst-margin	Mono frac
ℓ^2	$1.608 \times 10^4 \pm 1.0 \times 10^3$	0.000 ± 0.000	$2.505 \times 10^5 \pm 6.1 \times 10^4$	0.000 ± 0.000
ℓ^∞	$1.667 \times 10^4 \pm 1.7 \times 10^3$	0.000 ± 0.000	$2.628 \times 10^5 \pm 3.7 \times 10^4$	0.000 ± 0.000
graph	$\mathbf{1.315 \times 10^4 \pm 1.7 \times 10^3}$	0.000 ± 0.000	$\mathbf{2.918 \times 10^5 \pm 5.7 \times 10^4}$	0.000 ± 0.000
graph + structure	$2.514 \times 10^4 \pm 6.9 \times 10^2$	0.000 ± 0.000	$7.515 \times 10^4 \pm 1.4 \times 10^4$	0.000 ± 0.000

For A_1, the graph model achieves the best approximation accuracy across all four test metrics. The graph + structure model achieves the smallest mean monotonicity violation and fraction of violating pairs, as well as the largest worst monotonicity margin. Thus, the regularization can improve certain observed monotonicity metrics, although it does not uniformly improve all metrics.

For A_2, the graph model achieves the best reported approximation metrics on the high-frequency OOD test distribution. For all four models, no monotonicity violations are detected among the sampled pairs, as reflected by zero Mono mean-viol and Mono frac, together with positive Mono worst-margin. These empirical results, however, do not guarantee monotonicity of the learned operators beyond the sampled pairs.

D.2 Experiment for the set-valued operator

We consider the set-valued maximally monotone operator $A_3 = \partial\Phi$ on $L^2(0,1)$, where

$$\Phi(u) = \int_0^1 |u(t)|\,dt, \qquad A_3(u) = \big\{v \in L^2(0,1) : v(t) \in \partial|u(t)| \text{ for a.e. } t\big\},$$

and

$$\partial|x| = \begin{cases} \{-1\}, & x < 0, \\ [-1,1], & x = 0, \\ \{1\}, & x > 0. \end{cases} \tag{21}$$

This experiment separates the task of predicting a particular single-valued selector from the task of approximating the graph of a set-valued operator.

We use a periodic grid of $N_{\text{grid}} = 128$ points on $[0,1)$. An initial random function is generated as

$$\widetilde{u}(t) = a + \sum_{k=1}^{3} \frac{0.18}{k^{1.5}}\big(a_k \sin(2\pi k t) + b_k \cos(2\pi k t)\big), \tag{22}$$

where $a = 0.25\tanh Z$, $Z \sim \mathcal{N}(0,1)$, and $a_k, b_k \sim \mathcal{N}(0,1)$ independently. We then apply the numerical snapping rule

$$u(t) = \begin{cases} 0, & |\widetilde{u}(t)| \le \delta_{\text{snap}}, \\ \widetilde{u}(t), & |\widetilde{u}(t)| > \delta_{\text{snap}}, \end{cases} \qquad \delta_{\text{snap}} = 0.1. \tag{23}$$

The exact subdifferential $\partial|x|$ is set-valued only at $x = 0$. The snapping operation is therefore an explicit numerical device that creates zero-valued grid points; it is not a relaxation of the mathematical subdifferential. After snapping, we generate one admissible selector for each sample:

$$y(t) = \begin{cases} -1, & u(t) < 0, \\ \xi(t), & u(t) = 0, \\ 1, & u(t) > 0, \end{cases} \qquad \xi(t) \sim \text{Uniform}[-1,1], \tag{24}$$

with independent random choices at the zero-valued grid points. Thus, the generated pairs satisfy the pointwise graph relation $y(t) \in \partial|u(t)|$ on the discretization. The pointwise ℓ^2 and soft ℓ^∞ objectives compare predictions with this particular randomly generated selector. For A_3, the soft ℓ^∞ loss is implemented as the batch average of a soft maximum of the pointwise absolute error over the spatial grid.

Because the pointwise subdifferential is known explicitly, we use a problem-specific vertical graph residual

$$r(x,v) = \operatorname{dist}(v, \partial|x|) = \begin{cases} |v+1|, & x < 0, \\ \max\{|v|-1, 0\}, & x = 0, \\ |v-1|, & x > 0. \end{cases} \tag{25}$$

For a minibatch $\{u^{(n)}\}_{n=1}^B$ and the predictions $\widehat{A}(u^{(n)})$, the graph loss is

$$\mathcal{L}_{\text{graph}}^{A_3} = \frac{1}{B}\sum_{n=1}^{B} \tau_{\text{graph}} \log\left(\sum_{i=1}^{N_{\text{grid}}} \exp\left\{\frac{r(u_i^{(n)}, \widehat{A}(u^{(n)})_i)}{\tau_{\text{graph}}}\right\}\right), \qquad \tau_{\text{graph}} = 0.02. \tag{26}$$

Table 5: Hyperparameters for the set-valued experiment $A_3 = \partial\Phi$.

Parameter	Value
Grid points	128
Training samples / test samples	2000 / 400
Batch size	64
Fourier frequencies	$k = 1, 2, 3$
Bias scale / Fourier scale	0.25 / 0.18
Fourier decay exponent	1.5
Snapping threshold δ_{snap}	0.1
Selector at zero	$\text{Uniform}[-1, 1]$
Epochs / optimizer	80 / Adam
Learning rate / weight decay	5×10^{-4} / 10^{-6}
Gradient clipping threshold	1.0
Soft ℓ^∞ temperature	0.02
Graph temperature τ_{graph}	0.02
FNO modes / width / depth	16 / 64 / 2
Structured model parameter λ	0.01
Non-expansive penalty weight γ	0.01
Non-expansive pairs	512
Monotonicity evaluation pairs	128^2
Random seeds	$0, 1, 2$

Table 6: Test results for the set-valued operator $A_3 = \partial\Phi$.

Model	Test MSE	Mean Rel-ℓ^2	Worst Rel-ℓ^2
ℓ^2	$\mathbf{0.1022 \pm 0.0035}$	$\mathbf{0.3516 \pm 0.0080}$	$\mathbf{0.9070 \pm 0.1083}$
ℓ^∞	0.1561 ± 0.0144	0.4378 ± 0.0211	0.9305 ± 0.0827
graph	0.1096 ± 0.0049	0.3643 ± 0.0100	0.9300 ± 0.0893
graph + structure	0.1335 ± 0.0043	0.4031 ± 0.0074	0.9479 ± 0.0484

Model	Test Graph	Mono mean-viol	Mono worst-margin	Mono frac
ℓ^2	0.1147 ± 0.0017	$7.134 \times 10^{-8} \pm 1.2 \times 10^{-7}$	$-3.242 \times 10^{-4} \pm 5.3 \times 10^{-4}$	$1.628 \times 10^{-4} \pm 1.9 \times 10^{-4}$
ℓ^∞	0.5316 ± 0.0366	$2.952 \times 10^{-8} \pm 5.1 \times 10^{-8}$	$-1.665 \times 10^{-4} \pm 2.9 \times 10^{-4}$	$\mathbf{8.138 \times 10^{-5} \pm 1.4 \times 10^{-4}}$
graph	$\mathbf{0.0985 \pm 0.0002}$	$\mathbf{1.997 \times 10^{-9} \pm 2.5 \times 10^{-9}}$	$\mathbf{-1.636 \times 10^{-5} \pm 2.1 \times 10^{-5}}$	$\mathbf{8.138 \times 10^{-5} \pm 7.0 \times 10^{-5}}$
graph + structure	0.1030 ± 0.0016	$1.081 \times 10^{-8} \pm 1.0 \times 10^{-8}$	$-5.331 \times 10^{-5} \pm 4.9 \times 10^{-5}$	$2.035 \times 10^{-4} \pm 1.9 \times 10^{-4}$

In particular, when $u_i^{(n)} = 0$, every prediction in $[-1, 1]$ has zero pointwise residual regardless of the random label in (24). This differs from the pairwise soft graph loss (11) used for A_1 and A_2. The residual in (26) is not the theoretical local graph discrepancy of Definition 5, and no consistency theorem between these quantities is claimed.

The same four model types are compared as above. The graph + structure model combines (26) with the sampled non-expansiveness penalty, with weight $\gamma = 0.01$. As in the other experiments, the penalty encourages but does not guarantee global non-expansiveness.

Test MSE, Mean Rel-ℓ^2, and Worst Rel-ℓ^2 are computed against the sampled selector y in (24). For A_3, Test Graph is the average soft-max vertical residual (26) evaluated on the test data, not the pairwise metric (16). Monotonicity is evaluated on all ordered pairs of the first 128 test inputs, using the discrete L^2 inner product (grid average), the worst signed margin, and the violation fraction with threshold -10^{-8}. The two types of Test Graph values should therefore not be compared numerically across different target operators.

Table 5 summarizes the hyperparameter settings, while Table 6 reports the corresponding results.

For A_3, the ℓ^2 model achieves the best pointwise prediction accuracy, whereas the graph model achieves the smallest Test Graph error. The graph model also exhibits the smallest mean monotonicity violation and the largest worst monotonicity margin, although small violations are still detected. These

results illustrate the distinction between fitting a particular sampled selection and approximating the underlying set-valued graph.

More broadly, we note that constructing practical learning algorithms for genuinely set-valued targets is itself a nontrivial problem. The random-selection strategy adopted here provides one simple realization, but it does not amount to directly learning a set-valued output. Designing principled learning methods that directly represent and learn set-valued outputs remains an interesting topic for future research.