

GENERATIVE NEURAL OPERATORS OF LOG-COMPLEXITY CAN SIMULTANEOUSLY SOLVE INFINITELY MANY CONVEX PROGRAMS

ANASTASIS KRATSIOS*, ARIEL NEUFELD†, AND PHILIPP SCHMOCKER‡

Abstract. Neural operators (NOs) are a class of deep learning models designed to simultaneously solve infinitely many related problems by casting them into an infinite-dimensional space, whereon these NOs operate. A significant gap remains between theory and practice: worst-case parameter bounds from universal approximation theorems suggest that NOs may require an unrealistically large number of parameters to solve most operator learning problems, which stands in direct opposition to a slew of experimental evidence. This paper closes that gap for a specific class of NOs, generative equilibrium operators (GEOs), using (realistic) finite-dimensional deep equilibrium layers, when solving families of convex optimization problems over a separable Hilbert space X . Here, the inputs are smooth, convex loss functions on X , and outputs are the associated (approximate) solutions to the optimization problem defined by each input loss.

We show that when the input losses lie in suitable infinite-dimensional compact sets, our GEO can uniformly approximate the corresponding solutions to arbitrary precision, with rank, depth, and width growing only logarithmically in the reciprocal of the approximation error. We then validate both our theoretical results and the trainability of GEOs on three applications: (1) nonlinear PDEs, (2) stochastic optimal control problems, and (3) hedging problems in mathematical finance under liquidity constraints.

Keywords: Exponential Convergence, Proximal Splitting, Convex Optimization, Operator Learning, Stochastic Optimal Control, Non-Linear PDEs, Quadratic Hedging, Mathematical Finance.

1. Introduction. Neural operators (NOs) amortize the computational cost of solving large families of problems by learning reusable structure across *infinitely many* related tasks. Unfortunately, there is currently a large gap between NO theory and practice, since the approximation guarantees for neural operators suggest that, though NOs can approximately solve most infinite-dimensional problems [22, 30, 39, 40, 42, 58] they may need an exorbitant number of parameters [44, 45] to do so; unless the target operators is extremely smooth [2, 51]. This is surprising, as most operators encountered in practice are not that smooth; yet, there is a vast and well-documented literature showing that neural operators can successfully resolve most computational problems using a feasible number of parameters; e.g. [4, 10, 36, 41, 48, 54, 61, 70]. This large gap between theory and practice, thus cannot be resolved using tools from classical approximation theory.

This paper focuses precisely on closing this gap. We do so by 1) exhibiting a non-smooth but iterative structure which NOs can favourably leverage using their *depth*; and 2) tweaking standard NOs with deep equilibrium layers to provably take advantage of this structure and solve broad classes of infinite families of optimization problems with sub-linear parametric complexity. More precisely, we develop a NO solving (infinite) families of expressible as solutions to *convex* optimization problems of a “splittable” form

$$(1.1) \quad g \mapsto \operatorname{argmin}_{x \in X} \ell_{f,g}(x) \quad \text{with} \quad \ell_{f,g}(x) \stackrel{\text{def.}}{=} f(x) + g(x)$$

where X is a separable Hilbert space, $f : X \rightarrow (-\infty, \infty]$ is a proper, convex, and lower semicontinuous function, and $g : X \rightarrow \mathbb{R}$ is convex and Gâteaux differentiable with $(p-1)$ -Hölder continuous gradient for some $p \geq 2$. We approximate the associated *loss-to-solution* ($g \mapsto \text{minimizer of } f + g$, for fixed f) using neural operator-based foundation models for problems of the form (1.1) since they are core to a variety of scientific issues ranging from: parametric families of non-linear PDEs [8, 12, 14, 31, 35, 38, 38, 43, 46, 48, 50, 51, 52], stochastic optimal control [5, 9, 13, 19, 27, 34, 47, 67, 68], and quadratic hedging in mathematical finance [17, 32, 49, 56, 57, 62, 65, 66]. Additionally, any foundation model for the above can rapidly generate high-fidelity solutions that may either be used directly with minimal computational overhead or serve as inputs to a classical, case-specific downstream solver, yielding highly accurate solutions to a given convex optimization problem of the above form with little additional computational cost.

*Department of Mathematics, McMaster University and the Vector Institute, Canada (kratsioa@mcmaster.ca).

†Division of Mathematical Sciences, Nanyang Technological University, Singapore (ariel.neufeld@ntu.edu.sg).

‡Division of Mathematical Sciences, Nanyang Technological University, Singapore (philipp001@e.ntu.edu.sg).

Our solutions come in the form of a newly-designed variant on NOs using deep equilibrium (DE) layers, which is both non-deterministic, i.e. generative, and does not rely on infinite-dimensional DE layers (which, in general, need not be compatible with real-world computation). Our **Generative deep Equilibrium Operator** (GEO) architecture whose implicit bias encodes proximal forward-backward splitting procedures of [21] directly into its internal logic, allowing it to *simultaneously* solve infinite families of the convex optimization problems in (1.1) with minimal computational overhead. Our model leverages *proximal operators* as multivariate implicit nonlinear activation functions, thus extending standard **deep equilibrium** models (DEQ) [6] to infinite dimension, reflecting the recent developments in monotone DEQs [69], and which enjoy the convergence benefits of models leveraging fixed point iterations; e.g. DEQs with guarantees [29] in finite dimensions, or DEQs in infinite dimensions which either implicitly [28] or explicitly [26, 52] perform fixed point iterations. Additionally, the generative aspect of our neural operator model builds on the generative adversarial neural operators of [63] and allows for a greater diversity in its predictions through internal sources of randomness. Our generative DEQ lies at the intersection of *deep equilibrium* and *generative* modelling in *infinite dimensions*, specialized in *convex optimization* problems of a “splittable” form (1.1).

1.1. Main Contributions. Our main result (Theorem 3.2) shows that GEOs can approximate the *loss-to-solution* mapping of any admissible g in (1.1) for the corresponding splittable convex optimization problem over X . Critically, when the set of all admissible g is sufficiently well-behaved (formalized in (3.2)), the approximation can be achieved by GEOs whose depth grows at-most *logarithmically* in the reciprocal of the approximation error $\varepsilon > 0$, and whose width and rank do not grow exponentially therein. Moreover, if both f and all admissible g are Lipschitz with a shared worst-case Lipschitz constant, then our second main result (Theorem 3.3) shows that the optimal value itself can be recovered to roughly the same precision as the approximation accuracy of the loss-to-solution operator. Hence, *feasibly small* GEOs can approximately solve infinitely many (nonlinear) splittable convex optimization problems to high accuracy, thereby bypassing known limitations of general neural operator solutions when approximating arbitrary continuous or smooth solution operators [30, 45]. Our proof is based on the idea of approximately “unrolling” the proximal forward-backward splitting iterations of [21], which have recently found quantitative foundations in [16, 33], onto the layers of our neural operator architecture. Each of these results are predicated on the existence of a continuous approximate (η)-selectors for the coefficient (g) to solution operator for each splittable convex optimization problem in (1.1), with slack parameter $\eta > 0$ (Proposition 3.1).

1.2. Secondary Contributions. We then apply our main results to problems in non-linear partial differential equations (PDEs) (Section 4.1), stochastic optimal control (Section 4.2), and finally to optimal hedging in mathematical finance (Section 4.3). Each application explains and derives the relevant family of (non-linear) convex splittable optimization problems and is accompanied by a numerical illustration showing the reproducibility of our theoretical claims in each setting. An additional finite-dimensional application is included in our supplementary material (Section A.2).

1.3. Related Work. Our NO analysis resonates with recent efforts in scientific machine learning to embed algorithmic priors into learned architectures. These include PDE solvers using deep operator networks [48, 50, 52], neural realizations of classical schemes such as multi-grid or fixed-point constructions, e.g. Cauchy-Lipschitz, Lax-Milgram, Newton-Kantorovich theorems (see [26]), and the design of structured nonlinear mappings with guaranteed geometric convergence via nonlinear Perron-Frobenius theory [29]. We further highlight the connection between our neural operator architecture and the recent literature on operator learning in infinite dimensions [40, 46, 50] which typically suffers from the curse of dimensionality [45], algorithm unrolling [53, 55] which writes various forms of algorithmic logic directly into neural network layers and monotone operator theoretic perspectives on DEQ [7].

1.4. Organization of Paper. Section 2 compiles the preliminary background and notation required in the formulation of our main result and it introduces our GEO model. Section 3 contains the existence of

a continuous (approximate) loss-to-solution operator and our main approximation guarantees thereof. Section 4 contains worked out applications of our results to PDEs, stochastic optimal control, and mathematical finance. Section 5 contains a conclusion, whereas all proofs are relegated to Section 6. Additional background on proximal operators is included in our paper's supplementary material (see Appendix A).

2. Preliminaries. We now cover the background and terminology required to formulate our results.

Notation. Let $\mathbb{N} \stackrel{\text{def.}}{=} \{0, 1, 2, \dots\}$ and $\mathbb{N}_+ \stackrel{\text{def.}}{=} \{n \in \mathbb{N} : n > 0\}$. Given a vector field $V : \mathbb{R}^d \rightarrow \mathbb{R}^d$, we denote its support by $\text{supp}(V) \stackrel{\text{def.}}{=} \overline{\{x \in \mathbb{R}^d : V(x) \neq 0\}}$ where \bar{A} denotes the closure of a subset $A \subseteq \mathbb{R}^d$ in the norm topology. For each $N \in \mathbb{N}_+$, we define the N -simplex $\Delta_N \stackrel{\text{def.}}{=} \{w \in [0, 1]^N : \sum_{n=1}^N w_n = 1\}$. Let $\Gamma_0(X)$ denote the set of lower semi-continuous, proper, and convex (non-linear) maps from X to $(-\infty, \infty]$. We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all our random variables are defined.

For any $R \in \mathbb{N}_+$, we define the finite dimensional vector subspace $E_R \stackrel{\text{def.}}{=} \text{span}(\{e_j\}_{j=0}^{R-1}) \subseteq X$ and consider the *projection operator*

$$(2.1) \quad X \ni x \quad \mapsto \quad P_R(x) \stackrel{\text{def.}}{=} \sum_{j=0}^{R-1} \langle x, e_j \rangle e_j \in E_R,$$

the *lifting operator*

$$(2.2) \quad \mathbb{R}^R \ni z \stackrel{\text{def.}}{=} (z_0, \dots, z_{R-1})^\top \quad \mapsto \quad z^{\uparrow:R} \stackrel{\text{def.}}{=} \sum_{j=0}^{R-1} z_j e_j \in E_R,$$

and the *real-encoding operator*

$$(2.3) \quad X \ni x \quad \mapsto \quad x^{\downarrow:R} \stackrel{\text{def.}}{=} (\langle x, e_j \rangle)_{j=0}^{R-1} \in \mathbb{R}^R.$$

Observe that $(x^{\downarrow:R})^{\uparrow:R} = x$ for any $x \in E_R$ and $R \in \mathbb{N}_+$. Thus, in this sense, the operators $\cdot^{\uparrow:R}$ and $\cdot^{\downarrow:R}$ are purely formal identifications of E_R with \mathbb{R}^R and visa-versa.

The topology on Continuously Fréchet-Differentiable Operators. We henceforth equip $C(X, X)$ with the topology of uniform convergence on compact subsets of X . We equip $C^1(X, \mathbb{R})$ with the locally-convex topology τ generated by the family of semi-norms $\{p_K\}_K$ defined for any $g \in C^1(X, \mathbb{R})$ by

$$p_K(g) \stackrel{\text{def.}}{=} \sup_{x \in K} |g(x)| + \|\nabla g(x)\|_X$$

where the family $\{p_K\}_K$ is indexed over all non-empty compact subsets K of X . Note that, by construction, the locally-convex topology τ on $C^1(X, \mathbb{R})$ is not metrizable when X is not hemicompact; e.g. when X is a locally-compact metric space. Now, by definition of τ on $C^1(X, \mathbb{R})$ and the uniform convergence on compact sets, the topology on $C(X, X)$ guarantees the continuity of the following non-linear operator from $C^1(X, \mathbb{R})$ to $C(X, X)$ sending any $g \in C^1(X, \mathbb{R})$ to

$$(2.4) \quad C^1(X, \mathbb{R}) \ni g \quad \mapsto \quad \nabla g \in C(X, X).$$

Convex Analysis in Banach Spaces. The sub-differential of $f \in \Gamma_0(X)$ is defined as the set-valued mapping $\partial f : X \rightarrow X^*$ given for every $x \in X$ by

$$(2.5) \quad \partial f(x) = \{x^* \in X^* : \langle x^*, y - x \rangle \leq f(y) - f(x), \forall y \in X\}.$$

A point $\hat{x} \in X$ is a minimizer of f if and only if $0 \in \partial f(\hat{x})$. The sub-differential mapping $x \mapsto \partial f(x)$ has the property of monotonicity, i.e.,

$$(2.6) \quad \langle x_1^* - x_2^*, x_1 - x_2 \rangle \geq 0, \quad \forall x_1, x_2 \in X, \quad x_1^* \in \partial f(x_1), \quad x_2^* \in \partial f(x_2).$$

Let $g : X \rightarrow (-\infty, +\infty)$ be convex and Gâteaux differentiable with the gradient operator ∇g being $(p-1)$ -Hölder-continuous on X with $p \geq 2$, i.e., there exists a constant L such that:

$$(2.7) \quad \|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|^{p-1}, \quad \forall x, y \in X.$$

Our activation functions are defined using *proximal operators*, sometimes called the proximity operator, associated to any given $f \in \Gamma_0(X)$ by

$$(2.8) \quad \text{prox}_f(x) \stackrel{\text{def.}}{=} \underset{z \in X}{\text{argmin}} \ f(z) + \frac{1}{2}\|z - x\|_X^2$$

which is a well-defined Lipschitz (non-linear) monotone operator by; see e.g. [7, Chapter 24]. In the case where f is additionally Gâteaux differentiable, then we observe that

$$(2.9) \quad y = \text{prox}_f(x) \iff y = (\text{id}_X + \nabla f)^{-1}(x),$$

where $X \ni y \mapsto \nabla f(y) \in X$ is such that $\langle \nabla f(y), v \rangle_{X^* \times X} = Df(y)(v)$ for all $v \in X$, and where the notation $(\text{id}_X + \nabla f)^{-1}$ is defined in terms of a von Neumann series expansion. Henceforth X will be a separable *infinite-dimensional* Hilbert space with a distinguished *orthonormal basis* $(e_i)_{i=1}^\infty$.

2.1. Our Generative Equilibrium Operator. We would ideally like to use *deep equilibrium layers* to introduce nonlinearity into our neural operator, via the proximal operator $\text{prox}_f : X \rightarrow X$ associated to f (see (2.8)). In general, however, these operators may involve genuinely infinite-dimensional computations and thus may not be implementable on real-world machines. Using the projection operator P_R any $f \in \Gamma_0(X)$ defines a rank R multi-variate activation function $\sigma_f : X \rightarrow X$ sending any $x \in X$ to

$$(2.10) \quad \sigma_f(x) \stackrel{\text{def.}}{=} \sum_{j=0}^{R-1} \langle \text{prox}_f(x), e_j \rangle e_j.$$

If infinitely many parameters were processable on our idealized computer, then by setting $R = \infty$, the activation function σ_f would coincide with the proximal operator.

Importantly, unlike standard deep equilibrium layers for NOs, e.g. [52], the map σ_f is by construction *implementable* using on a *finitely parameterized* subspace of X ; which need not be true for the proximal operator (equilibrium layer) in (2.8). Independently of the *generative* aspect of our neural operator, our model diverges from the standard NO build in a number of subtle but key ways. Most strikingly, we do not leverage a univariate activation, acting pointwise, but rather a *structurally-dependent* multivariate activation function. For every problem 1.1, the (potentially) non-differentiable component of the objective function, namely f , includes a finite-rank operator which introduces non-linearity into our neural operator's updates.

We additionally incorporate a gated residual connection, which allows information to be passed forward following the non-linear processing occurring at each layer. At first glance, this is motivated by the empirically [15] and theoretically observed loss-landscape regularization effects of residual connections [64]. However, as we will see in the proofs section, the connection runs deeper in our setting in connection with Forward-backward proximal splitting algorithms [21].

DEFINITION 2.1 (Generative Equilibrium Operator). Fix a rank $R \in \mathbb{N}_+$, a sampling level $M \in \mathbb{N}_+$, a depth $L \in \mathbb{N}_+$, a source of noise $\xi \in L^1(E_R)$, and some $f \in \Gamma_0(X)$. Then, a *Generative Equilibrium Operator* with activation function σ_f is a map $\mathcal{G} : \Omega \times C(X) \rightarrow E_R$ given for any $x \in X$ by

$$\mathcal{G}(\omega, g) \stackrel{\text{def.}}{=} \left(A^{(L+1)} x^{(L+1) \downarrow : M} \right)^{\uparrow : M},$$

and iteratively for $l = 0, \dots, L + 1$ via

$$x^{(l+1)} \stackrel{\text{def.}}{=} \underbrace{\gamma^{(l)} x^{(l)}}_{\text{Skip Connection}} + \overbrace{(1 - \gamma^{(l)})}^{\text{Gating}} \sigma_f \left(\underbrace{A^{(l)} x^{(l)}}_{\text{Weights}} + \underbrace{\left[B^{(l)} \left(\underbrace{g(x^{(l)} + x_m^{(l)})}_{\text{Adaptive Sampling}} \right)^M_{m=1} + \underbrace{b^{(l)}}_{\text{Bias}} \right]^{\uparrow:M}}_{\text{g-Dependent weights}} \right),$$

$$x^{(0)} \stackrel{\text{def.}}{=} \xi(\omega),$$

where $A^{(l)} \in \mathbb{R}^{R \times R}$ are weight matrices, $B^{(l)} \in \mathbb{R} \times \mathbb{M}$ are weight matrices, and $b^{(l)} \in \mathbb{R}^R$ are bias vectors, $\{x_m^{(l)}\}_{m,l=0}^{M,L} \subset X$ are sample points, and $\gamma^{(l)} \in [0, 1]$ are gating coefficients, $l = 0, \dots, L$.

3. Main Guarantee. We begin by establishing the existence of a (nonlinear), continuous, approximate optimal selection operator, which we aim to approximate using our Generative Equilibrium Operator for (1.1). Note that, in general, a continuous optimal selector (corresponding to $\eta = 0$) may not exist. Moreover, even if a Borel-measurable selector does exist, it typically cannot be approximated by continuous objects such as our Generative Equilibrium Operator.

Since we are only implementing an approximate solution operator, an approximation error is inevitable. Consequently, there is no issue in introducing an additional—but arbitrarily small—sub-optimality error in the solution operator in exchange for continuity, and hence, approximability. Of course, both sources of error can be asymptotically driven to zero, as is standard in approximation theory.

Importantly, the near optimality is independent of the input in the class \mathcal{X}_λ of inputs $g \in C^1(X)$ with uniformly bounded Fréchet gradient defined by

$$(3.1) \quad \mathcal{X}_\lambda \stackrel{\text{def.}}{=} \left\{ g \in C^1(X) : \nabla g \text{ is convex and } \lambda\text{-Lipschitz} \right\}.$$

PROPOSITION 3.1 (Existence of a randomized $\mathcal{O}(\eta)$ -optimal selector). *For every approximate solution parameter $\eta > 0$ and any $\lambda > 0$ there exists an “approximation solution” operator $S_\eta : \Omega \times C^1(X) \rightarrow X$ satisfying: for every $\omega \in \Omega$, $S_\eta(\omega, \cdot) : C^1(X) \rightarrow X$ is continuous and for every $g \in \mathcal{X}_\lambda$ it holds that*

$$\ell_{f,g}(S_\eta(\omega, g)) - \inf_{x \in X} \ell_{f,g}(x) \lesssim_\omega \eta,$$

where $\ell_{f,g}$ is defined in (1.1) and \lesssim_ω hides a multiplicative constant depending only on ω and on λ (thus independent of g and of η).

For any $r, \lambda \geq 0$, we consider the functions in $\mathcal{X}_\lambda(r)$ whose Fréchet gradient is well-explained with few basis factors. More precisely,

$$(3.2) \quad \mathcal{X}_\lambda(r) \stackrel{\text{def.}}{=} \left\{ g \in \mathcal{X}_\lambda : \sum_{i=R}^{\infty} |(\partial_t g(x + te_i))|_{t=0}|^2 \leq r 2^{-R/2} \text{ for all } x \in X \text{ and } R \in \mathbb{N} \right\}.$$

The set $\mathcal{X}_\lambda(r)$ are a take on the exponentially ellipsoidal sets of [3, 30], which abstract the Fourier analytic characterization of smooth and rapidly decaying functions [60] where the rapid decay conditions are on the function’s Fréchet gradient and not on the function itself. Now that we know there exists a well-posed, continuous $\mathcal{O}(\eta)$ -optimal solution operator for the family of convex optimization problems in (1.1), indexed by $g \in \mathcal{X}_\lambda(r)$ for any given $\eta > 0$ and $r, \lambda > 0$, we can meaningfully consider approximating them.

THEOREM 3.2. *For any $r, \lambda \geq 0$, and $f \in \Gamma_0(X)$, and any approximation error $\varepsilon > 0$ there is a Generative Equilibrium Operator of rank $R \in \mathcal{O}(\log(1/\varepsilon))$, depth $L \in \mathcal{O}(\log(1/\varepsilon))$, and with $M \in \mathcal{O}(\log(1/\varepsilon))$ sample points satisfying*

$$(3.3) \quad \sup_{g \in \mathcal{X}_\lambda(r)} \|S_\eta(\omega, g) - \mathcal{G}(\omega, g)\|_X \lesssim_\omega \varepsilon \quad \mathbb{P}\text{-a.s.}$$

where \lesssim_ω hides a multiplicative constant depending only on the draw of $\omega \in \Omega$ and is independent of η , ε , and of any $g \in \mathcal{X}_\lambda(r)$.

If the function f in (1.1) is Lipschitz, then the Generative Equilibrium Operator \mathcal{G} from Theorem 3.2 approximately solves the splitting problem in (1.1) for any suitably regular input g .

THEOREM 3.3 (Simultaneous Approximately Optimal Splitting). *Fix $\lambda_f, \lambda_g, \lambda \geq 0$, consider the setting of Theorem 3.3, and let \mathcal{G} be a GEO satisfying (3.3). If f is additionally λ_f -Lipschitz, then for any $g \in \mathcal{X}_\lambda(r)$ with λ -Lipschitz Fréchet gradient we additionally have*

$$(3.4) \quad \ell_{f,g}(\mathcal{G}(g)) - \inf_{x \in X} \ell_{f,g}(x) \lesssim_\omega \varepsilon + \eta$$

where \lesssim_ω hides a constant depending only on the draw of $\omega \in \Omega$ and is independent of η , ε , and of any $g \in \mathcal{X}_\lambda(r)$.

Why Approximate the η -Solution Operator Instead of the True Solution Operator? A subtle but important point is that the continuity of the η -approximate solution operator S_η allows it to be approximated by continuous objects, such as our GEO models in Theorem 3.2, even when the true solution operator may not be approximable in this way. Crucially, since S_η always achieves an η -optimal loss and is continuous, it admits such approximations with only an additional additive error of at most η in the final loss (Theorem 3.3). Note that η may be chosen arbitrarily small.

4. Numerical Experiments. We illustrate in four different numerical examples how Generative Equilibrium Operators can be implemented on a computer to learn convex splitting problems of the form (1.1)¹.

4.1. Learning the solution of a parametric family of non-linear PDEs. Before applying the forward-backward proximal splitting algorithm to non-linear partial differential equations (PDEs), we first recall that the proximal operator can be understood as implicit Euler discretization of a gradient flow differential equation. More precisely, for a Hilbert space X and a proper, lower semicontinuous, and convex function $f : X \rightarrow (-\infty, \infty]$, we consider the differential

$$(4.1) \quad \partial_t y(t) \in -\partial f(y(t)), \quad t \in [0, \infty).$$

The solution $y : [0, \infty) \rightarrow X$ of (4.1) is called the gradient flow of $f : X \rightarrow (-\infty, \infty]$. If $f : X \rightarrow (-\infty, \infty]$ is differentiable, then an implicit Euler discretization of (4.1) along a partition $0 < t_0 < t_1 < \dots$ leads to

$$\frac{y(t_{k+1}) - y(t_k)}{t_{k+1} - t_k} \approx -\nabla f(y(t_{k+1})), \quad k \in \mathbb{N}.$$

Hence, we observe that

$$y(t_{k+1}) = (\text{id}_X + (t_{k+1} - t_k)\nabla f)^{-1}(y(t_k)) = \text{prox}_{(t_{k+1} - t_k)f}(y(t_k)), \quad k \in \mathbb{N}.$$

Thus, the proximal minimization algorithm coincides with the implicit Euler method for numerically solving the gradient flow differential equation (4.1). Now, for $X \stackrel{\text{def.}}{=} L^2(U) \stackrel{\text{def.}}{=} L^2(U, \mathcal{L}(U), du)$ with $U \subseteq \mathbb{R}^d$ and a given initial condition $y_0 \in L^2(U)$, we consider a non-linear partial differential equation (PDE) of the form

$$(4.2) \quad \frac{\partial y}{\partial t}(t, u) + (\mathcal{A}^* \mathcal{A}y(t, \cdot))(u) + q(y(t, u)) = 0, \quad (t, u) \in (0, T) \times U,$$

¹All numerical experiments have been implemented in Python using the Tensorflow package and were executed on a high-performing computing (HPC) cluster provided by the Digital Research Alliance of Canada. The code can be found under the following link: <https://github.com/psc25/GenerativeEquilibriumOperator>.

with initial condition $y(0, u) = y_0(u)$, $u \in U$, where $\mathcal{A} : \text{dom}(\mathcal{A}) \subseteq L^2(U) \rightarrow L^2(U)$ is a (possibly unbounded) linear operator² with adjoint \mathcal{A}^* , and where $q : \mathbb{R} \rightarrow \mathbb{R}$ is a continuous non-linear function with $q \circ x \in L^2(U)$ for all $x \in L^2(U)$, whose antiderivative $Q : \mathbb{R} \rightarrow \mathbb{R}$ is convex and satisfies $Q \circ x \in L^1(U)$ for all $x \in L^2(U)$. Then, by applying an explicit Euler step to $\mathcal{A}^* \mathcal{A} y(t_k, \cdot)$ and an implicit Euler step to $q \circ y(t_k, \cdot)$ along a partition $0 < t_0 < t_1 < \dots$, we obtain that

$$\frac{y(t_{k+1}, \cdot) - y(t_k, \cdot)}{t_{k+1} - t_k} \approx -\mathcal{A}^* \mathcal{A} y(t_k, \cdot) - q(y(t_{k+1}, \cdot)),$$

which is known as forward-backward splitting of PDEs (see also [8, 59]). Moreover, we define the function $f(x) = \int_U Q(x(u)) du$ and $g(x) \stackrel{\text{def}}{=} \frac{1}{2} \|\mathcal{A}x\|_{L^2(U)}^2$ satisfying for every $v \in L^2(U)$ that

$$Df(x)(v) = \frac{d}{dh} \Big|_{h=0} \left(\int_U Q((x + hv)(u)) du \right) = \int_U q(x(u)) v(u) du = \langle q \circ x, v \rangle_{L^2(U)}$$

and

$$Dg(x)(v) = \frac{d}{dh} \Big|_{h=0} \left(\frac{1}{2} \|\mathcal{A}(x + hv)\|_{L^2(U)}^2 \right) = \langle \mathcal{A}x, \mathcal{A}v \rangle_{L^2(U)} = \langle \mathcal{A}^* \mathcal{A}x, v \rangle_{L^2(U)},$$

which shows that $\nabla f(x) = q \circ x$ and $\nabla g(x) = \mathcal{A}^* \mathcal{A}x$. Hence, by using (2.9), we observe that

$$\begin{aligned} y(t_{k+1}, \cdot) &= (\text{id}_X + (t_{k+1} - t_k) \nabla f)^{-1} (\text{id}_X - (t_{k+1} - t_k) \mathcal{A}^* \mathcal{A}) y(t_k, \cdot) \\ &= \text{prox}_{(t_{k+1} - t_k) f} (y(t_k, \cdot) - (t_{k+1} - t_k) \mathcal{A}^* \mathcal{A} y(t_k, \cdot)), \end{aligned}$$

which shows that the proximal operator can be applied to learn the solution of the PDE (4.2).

EXAMPLE 4.1. For the Hilbert space $X \stackrel{\text{def}}{=} L^2(\mathbb{R}) \stackrel{\text{def}}{=} L^2(\mathbb{R}, \mathcal{L}(\mathbb{R}), du)$ and $T, \nu > 0$, we consider the PDE (4.2) of linear reaction–diffusion type with $\mathcal{A}^* \mathcal{A}x \stackrel{\text{def}}{=} -\nu x''$ and $q(x(u)) = \min(x(u), 0)$, i.e.

$$(4.3) \quad \frac{\partial y}{\partial t}(t, u) - \nu \frac{\partial^2 y}{\partial u^2}(t, u) + \frac{1}{2} \min(y(t, u), 0) = 0, \quad (t, u) \in (0, T) \times \mathbb{R},$$

with initial condition $y_0(y) = 5u e^{-u^2}$, $u \in U$, where $\mathcal{A}x \stackrel{\text{def}}{=} \sqrt{\nu} x'$ with $\mathcal{A}^* x = -\mathcal{A}x$ due to integration by parts, and where $Q(s) \stackrel{\text{def}}{=} \mathbf{1}_{(-\infty, 0)}(s) \frac{s^2}{4}$. The proximal operator of $f(x) \stackrel{\text{def}}{=} \int_{\mathbb{R}} Q(x(u)) du$ is given by $\text{prox}_f(x) = (u \mapsto x(u) - \frac{1}{4} \min(x(u), 0))$, whereas $g_\nu(x) \stackrel{\text{def}}{=} \frac{1}{2} \|\mathcal{A}x\|_{L^2(\mathbb{R})}^2 = \frac{\nu}{2} \|x'\|_{L^2(\mathbb{R})}^2$. In this setting, we aim to learn the operator

$$(4.4) \quad \mathbb{R} \ni \nu \mapsto \mathcal{S}(g_\nu) \stackrel{\text{def}}{=} y(T, \cdot) = \arg \min_{x \in X} (f(x) + g_\nu(x)) \in L^2(\mathbb{R}),$$

by a Generative Equilibrium Operator \mathcal{G} of rank $R = 8$, depth $L = 10$, and width $M = 20$. To this end, we choose the Hermite-Gaussian functions $(e_j)_{j \in \mathbb{N}}$ as basis of $L^2(\mathbb{R})$, which are defined by $e_j(u) \stackrel{\text{def}}{=} \frac{H_j(u)}{(2^j j!)^{1/2}} \frac{e^{-u^2/2}}{\pi^{1/4}}$, for $u \in \mathbb{R}$, where $(H_j)_{j \in \mathbb{N}}$ are the physicist's Hermite polynomials (see [1, Equation 22.2.14]). Moreover, we apply the Adam algorithm over 20000 epochs with learning rate 10^{-4} to train the Generative Equilibrium Operator on a training set consisting of 400 randomly initialized parameters $\nu_1, \dots, \nu_{400} \in [0.01, 0.4]$. In addition, we evaluate its generalization performance every 250-th epoch on a test set consisting of 100 randomly initialized parameters $\nu_{401}, \dots, \nu_{500} \in [0.01, 0.4]$. Hereby, the reference solution $y(T, \cdot) \stackrel{\text{def}}{=} \mathcal{S}(g_\nu)$ of the non-linear PDE (4.3) is approximated by using a Multilevel Picard (MLP) algorithm (see, e.g., [24, 25]). The results are reported in Figure 4.1.

²However, this is not an issue as we can simply consider a bounded linear extension thereof by the Benyamini-Lindenstrauss theorem; see e.g. [11, Theorem 1.12]; which we may somewhat abusively also denote by \mathcal{A} .

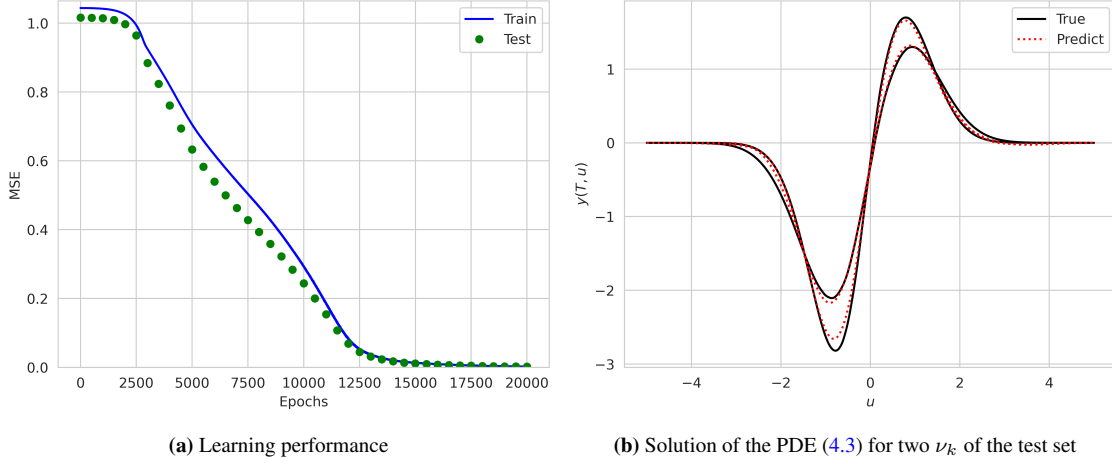


Fig. 4.1: Learning the map (4.4) returning the solution of the parametric PDE (4.3) by a Generative Equilibrium Operator \mathcal{G} . In (a), the learning performance is displayed in terms of the mean squared error (MSE) $\frac{1}{|K|} \sum_{k \in K} \|\mathcal{S}(g_{\nu_k}) - \mathcal{G}(g_{\nu_k})\|^2$ on the training set (label “Train”) and test set (label “Test”). In (b), the predicted solution $\mathcal{G}(g_{\nu_k})$ (label “Predict”) is compared to the true solution $y(T, \cdot) = \mathcal{S}(g_{\nu_k})$ (label “True”) for two ν_k of the test set.

4.2. Stochastic optimal control. In this section, we apply the proximal learning framework to solve the stochastic optimal control problem. For $T > 0$, a filtered probability space $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{P})$ with filtration $\mathbb{F} \stackrel{\text{def.}}{=} (\mathcal{F}_t)_{t \in [0, T]}$ satisfying the usual conditions, and an \mathbb{F} -adapted processes $x : [0, T] \times \Omega \rightarrow \mathbb{R}^n$ with $\mathbb{E}[\int_0^T \|x_t\|^2 dt] < \infty$, we assume that $y : [0, T] \times \Omega \rightarrow \mathbb{R}^d$ is a unique strong solution of the SDE

$$dy_t = \mu(t, y_t, x_t)dt + \sigma(t, y_t, x_t)dW_t, \quad t \in [0, T],$$

where $y_0 \in \mathbb{R}^d$, $\mu : [0, T] \times \mathbb{R}^d \times \mathbb{R}^n \rightarrow \mathbb{R}^d$ and $\sigma : [0, T] \times \mathbb{R}^d \times \mathbb{R}^n \rightarrow \mathbb{R}^{d \times d}$ are sufficiently regular functions, and where W is a d -dimensional Brownian motion. We denote by X the Hilbert space of \mathbb{F} -adapted processes $x : [0, T] \times \Omega \rightarrow \mathbb{R}^n$ with $\|x\|_X \stackrel{\text{def.}}{=} \mathbb{E}[\int_0^T \|x_t\|^2 dt] < \infty$. Besides using $f : X \rightarrow (-\infty, \infty]$ to implement some constraints, we minimize the objective function $g : X \rightarrow \mathbb{R}$ given by

$$g(x) = \mathbb{E} \left[\int_0^T (-c)(t, y_t, x_t)dt + (-u)(y_T) \right],$$

which is equivalent to expected utility maximization from consumption (with function $c : [0, T] \times \mathbb{R}^d \times \mathbb{R}^n \rightarrow \mathbb{R}$) and from terminal wealth (with function $u : \mathbb{R}^d \rightarrow \mathbb{R}$). Under regularity assumptions on μ , σ , c , and u , the corresponding value function satisfies a Hamilton-Jacobi-Bellman equation (see [27, 67, 68] for details).

EXAMPLE 4.2. We consider Merton’s optimal investment problem over a finite time horizon $T > 0$. For $r, \sigma > 0$ and $\mu \in \mathbb{R}$, we model the stock price S and the “risk-less” bond price B by

$$dS_t = S_t(\mu dt + \sigma dW_t), \quad dB_t = B_t r dt, \quad t \in [0, T],$$

with initial values $S_0 > 0$ and $B_0 = 1$, where W is a d -dimensional Brownian motion. Moreover, we assume that \mathbb{F} is the \mathbb{P} -completion of the filtration generated by W . In this case, if $x \stackrel{\text{def.}}{=} (x_t)_{t \in [0, T]}$ denotes the value

invested into the stock, then the wealth process $y \stackrel{\text{def.}}{=} (y_t)_{t \in [0, T]}$ of the corresponding self-financing trading strategy satisfies

$$\frac{y_t}{B_t} = y_0 + \int_0^t \frac{x_s}{B_s} ((\mu - r)ds + \sigma dW_s), \quad t \in [0, T],$$

for some initial value $y_0 \stackrel{\text{def.}}{=} 1$ (see [37, Equation 3.1]). In addition, we define the objective functions

$$f(x) = \begin{cases} 0, & \text{if } x_t(\omega) \in [0, \infty) \text{ for all } (t, \omega) \in [0, T] \times \Omega \\ \infty, & \text{otherwise,} \end{cases} \quad g_u(x) = \mathbb{E} [(-u)(y_T)],$$

which corresponds to utility maximization from terminal wealth (with utility function u) under the constraint that $x_t(\omega) \in [0, \infty)$ for all $(t, \omega) \in [0, T] \times \Omega$. The proximal operator of $f : X \rightarrow (-\infty, \infty]$ is given by

$$\text{prox}_f(x) = \text{proj}_{[0, \infty)}(x(\cdot)),$$

for all $x \in X$. Furthermore, by using the market price of risk $\lambda \stackrel{\text{def.}}{=} \frac{\mu - r}{\sigma} \in \mathbb{R}$, we define the process $Z_t \stackrel{\text{def.}}{=} \exp(-\lambda W_t - \frac{\lambda^2}{2}t)$, $t \in [0, T]$ and conclude that $(S_t/B_t)_{t \in [0, T]}$ is by Girsanov's theorem a martingale under the equivalent measure $\mathbb{Q} \sim \mathbb{P}$ with density $\frac{d\mathbb{Q}}{d\mathbb{P}} \stackrel{\text{def.}}{=} Z_T$. We can apply [37, Theorem 3.7.6] to obtain that the optimal portfolio $x \stackrel{\text{def.}}{=} (x_t)_{t \in [0, T]}$ with respect to the above utility maximization problem is given by

$$x_t = \frac{\psi_t}{\sigma H_t} + \frac{\lambda}{\sigma H_t} \mathbb{E}[H_T \xi | \mathcal{F}_t], \quad t \in [0, T],$$

where $H \stackrel{\text{def.}}{=} (H_t)_{t \in [0, T]} \stackrel{\text{def.}}{=} (Z_t/B_t)_{t \in [0, T]}$, where $\xi \stackrel{\text{def.}}{=} I(\mathcal{Y}(y_0)H_T)$ with I being a left-inverse of u' and \mathcal{Y} being a right-inverse of $\mathcal{X}(y) \stackrel{\text{def.}}{=} \mathbb{E}[H_T I(yH_T)]$, and where $\psi \stackrel{\text{def.}}{=} (\psi_t)_{t \in [0, T]}$ satisfies $y_0 + \int_0^t \psi_s dW_s = \mathbb{E}[H_T \xi | \mathcal{F}_t]$ for all $t \in [0, T]$. In Table 4.1, we compute the optimal portfolios for some utility functions.

$u(x)$	$I(y)$	$\mathcal{X}(y)$	$\mathcal{Y}(y_0)$	ξ
$\begin{cases} \frac{(x-x_0)^{1-\eta}}{1-\eta}, & \eta \neq 1, \\ \ln(x-x_0), & \eta = 1. \end{cases}$	$\frac{1}{y^{1/\eta}} + x_0$	$\frac{1}{y^{1/\eta}} \mathbb{E}[H_T^{1-1/\eta}] - \frac{x_0}{B_T}$	$\frac{\mathbb{E}[H_T^{1-1/\eta}]^\eta}{(y_0-x_0/B_T)^\eta}$	$\frac{y_0-x_0/B_T}{\mathbb{E}[H_T^{1-1/\eta}]H_T^{1/\eta}} + x_0$
$\mathbb{E}[H_T \xi \mathcal{F}_t]$	ψ_t	x_t		
$\left(y_0 - \frac{x_0}{B_T}\right) \frac{Z_t^{1-1/\eta}}{\exp\left(\frac{\lambda^2}{2} \frac{1-\eta}{\eta^2} t\right)} + \frac{x_0 Z_t}{B_T}$	$-\lambda \left(y_0 - \frac{x_0}{B_T}\right) \frac{(1-1/\eta) Z_t^{1-1/\eta}}{\exp\left(\frac{\lambda^2}{2} \frac{1-\eta}{\eta^2} t\right)} - \lambda \frac{x_0 Z_t}{B_T}$	$\frac{\mu-r}{\sigma^2 \eta} \frac{(y_0-x_0/B_T) Z_t^{1-1/\eta}}{H_t \exp\left(\frac{\lambda^2}{2} \frac{1-\eta}{\eta^2} t\right)}$		

Table 4.1: Computation of optimal portfolios for power and logarithmic utility functions given by $u(x) \stackrel{\text{def.}}{=} \frac{(x-x_0)^{1-\eta}}{1-\eta}$ if $\eta \in (0, \infty) \setminus \{1\}$ and $u(x) \stackrel{\text{def.}}{=} \ln(x-x_0)$ if $\eta = 1$, where $x_0 \in \mathbb{R}$ is the reference point.

Now, we consider the closed vector subspace $X \subseteq L^2([0, T] \times \Omega, \mathcal{B}([0, T]) \otimes \mathcal{A}, dt \otimes d\mathbb{P})$ of W -Markovian processes, i.e. \mathbb{F} -predictable processes $x \stackrel{\text{def.}}{=} (x_t)_{t \in [0, T]}$ such that x_t is $\sigma(W_t)$ -measurable, for all $t \in [0, T]$. In this setting, we learn the solution operator of the utility maximization problem

$$(4.5) \quad \{u : D \subseteq \mathbb{R} \rightarrow \mathbb{R} \text{ is concave}\} \ni u \quad \mapsto \quad \mathcal{S}(g_u) \stackrel{\text{def.}}{=} \arg \min_{\substack{x \in X \\ x_t \geq 0}} \mathbb{E} [(-u)(y_T)] = \arg \max_{\substack{x \in X \\ x_t \geq 0}} \mathbb{E} [u(y_T)] \in X$$

by a Generative Equilibrium Operator of rank $R = 10$, depth $L = 10$, and width $M = 40$. To this end, we choose the non-orthogonal basis $(e_{j_1, j_2})_{j_1, j_2 \in \mathbb{N}}$ of X defined by $e_{j_1, j_2}(t) \stackrel{\text{def.}}{=} t^{j_1} W_t^{j_2}$, $t \in [0, T]$,

whence the coefficients of any $x \in X$ with respect to $(e_{j_1, j_2})_{j_1, j_2 \in \mathbb{N}}$ can be computed with the help of the Gram matrix (see the code). Moreover, we apply the Adam algorithm over 5000 epochs with learning rate $5 \cdot 10^{-5}$ and batchsize 100 to train the Generative Equilibrium Operator on a training set consisting of 400 utility functions $u_k(y) \stackrel{\text{def.}}{=} \frac{(y-x_{0,k})^{1-\eta_k}}{1-\eta_k}$, $k = 1, \dots, 400$. In addition, we evaluate its generalization performance every 125-th epoch on a test set consisting of 100 utility functions $u_k(y) \stackrel{\text{def.}}{=} \frac{(y-x_{0,k})^{1-\eta_k}}{1-\eta_k}$, $k = 401, \dots, 500$. Hereby, the risk aversion parameters $\eta_1, \dots, \eta_{500} \in [0.25, 0.75)$ and the reference points $x_{0,1}, \dots, x_{0,500} \in [0, \infty)$ are randomly initialized. The results are reported in Figure 4.2.

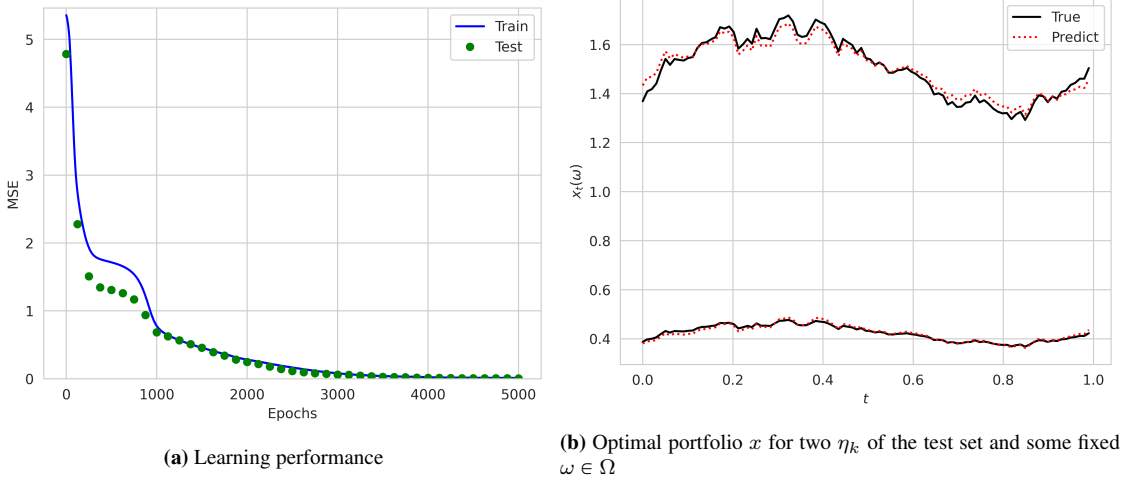


Fig. 4.2: Learning the solution operator \mathcal{S} of the utility maximization problem (4.5) by a Generative Equilibrium Operator \mathcal{G} . In (a), the learning performance is displayed in terms of the mean squared error (MSE) $\frac{1}{|K|} \sum_{k \in K} \|\mathcal{S}(g_{u_k}) - \mathcal{G}(g_{u_k})\|^2$ on the training set (label “Train”) and test set (label “Test”). In (b), the predicted solution $\mathcal{G}(g_{u_k})$ (label “Predict”) is compared to the true solution $\mathcal{S}(g_{u_k})$ (label “True”) for two η_k of the test set and some fixed $\omega \in \Omega$.

4.3. Quadratic hedging with liquidity constraint. In an incomplete financial market model, we learn the pricing/hedging operator that returns for a given financial derivative an approximation of the optimal hedging strategy in the sense of quadratic hedging under an additional liquidity constraint. For $T > 0$, a filtered probability space $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{P})$ with filtration $\mathbb{F} \stackrel{\text{def.}}{=} (\mathcal{F}_t)_{t \in [0, T]}$ satisfying the usual conditions, and a continuous strictly positive semimartingale $S \stackrel{\text{def.}}{=} (S_t)_{t \in [0, T]}$ with decomposition $S_t = S_0 + A_t + M_t$, $t \in [0, T]$, into a process of finite variation $A \stackrel{\text{def.}}{=} (A_t)_{t \in [0, T]}$ and a local martingale $M \stackrel{\text{def.}}{=} (M_t)_{t \in [0, T]}$ with $A_0 = M_0 = 0$, we consider the Hilbert space $\mathbb{R} \oplus L^2(S)$ of tuples $(x, \theta) \in \mathbb{R} \oplus L^2(S)$ equipped with the inner product $\langle (x, \theta), (y, \vartheta) \rangle_{\mathbb{R} \oplus L^2(S)} = xy + \langle \theta, \vartheta \rangle_{L^2(S)}$, where $L^2(S)$ denotes the Hilbert space of \mathbb{F} -predictable processes $\theta \stackrel{\text{def.}}{=} (\theta_t)_{t \in [0, T]}$ such that $\mathbb{E}[(\int_0^T |\theta_t dA_t|)^2]^{1/2} + \mathbb{E}[\int_0^T \theta_t^2 d\langle M \rangle_t]^{1/2} < \infty$, equipped with the inner product $\langle \theta, \vartheta \rangle_{L^2(S)} = \mathbb{E}[(\int_0^T |\theta_t dA_t|)(\int_0^T |\vartheta_t dA_t|)] + \mathbb{E}[\int_0^T \theta_t \vartheta_t d\langle M \rangle_t]$. For a given liquidity constraint $C > 0$ and a financial derivative $H \in L^2(\mathbb{P})$, we aim to minimize the hedging error

$$\inf_{\substack{(x, \theta) \in \mathbb{R} \oplus L^2(S) \\ x \in [0, C]}} \mathbb{E} \left[\left(H - x - \int_0^T \theta_t dS_t \right)^2 \right] = \inf_{(x, \theta) \in \mathbb{R} \oplus L^2(S)} (f(x, \theta) + g_H(x, \theta)),$$

where $f(x, \theta) \stackrel{\text{def}}{=} 0$ if $x \in [0, C]$, and $f(x, \theta) \stackrel{\text{def}}{=} \infty$ otherwise, and $g_H(x, \theta) \stackrel{\text{def}}{=} \mathbb{E}[(H - x - \int_0^T \theta_t dS_t)^2]$. Hereby, we observe that the proximal operator of $f : \mathbb{R} \oplus L^2(S) \rightarrow (-\infty, \infty]$ is given by

$$\text{prox}_f(x, \theta) = \arg \min_{(y, \vartheta) \in \mathbb{R} \oplus L^2(S)} \left(f(y, \vartheta) + \|(y, \vartheta) - (x, \theta)\|_{\mathbb{R} \oplus L^2(S)}^2 \right) = \left(\text{proj}_{[0, C]}(x), (\theta_t)_{t \in [0, T]} \right),$$

where $\text{proj}_{[0, C]}(s) \stackrel{\text{def}}{=} \arg \min_{t \in [0, C]} |s - t| = \max(\min(s, C), 0)$.

EXAMPLE 4.3. We consider the Heston model with stock price $S \stackrel{\text{def}}{=} (S_t)_{t \in [0, T]}$ and stochastic volatility $V \stackrel{\text{def}}{=} (V_t)_{t \in [0, T]}$ following the SDEs

$$\begin{aligned} dS_t &= \sqrt{V_t} S_t dW_t^1, \\ dV_t &= \kappa(\theta - V_t)dt + \sigma \sqrt{V_t} dW_t^2, \end{aligned}$$

where $\kappa, \theta, \sigma > 0$ and $d\langle W^1, W^2 \rangle_t = \rho dt$ for some $\rho \in [-1, 1]$. Moreover, we assume that \mathbb{F} is the \mathbb{P} -completion of the filtration generated by W^1 and W^2 . Hereby, we restrict ourselves to the closed vector subspace $X \subseteq \mathbb{R} \oplus L^2(S)$ of tuples $(x, \theta) \in \mathbb{R} \oplus L^2(S)$ such that θ is (S, V) -Markovian in the sense that x_t is $\sigma(S_t, V_t)$ -measurable, for all $t \in [0, T]$. In this setting, we learn the operator returning the price and optimal trading strategy in the sense of quadratic hedging by a Generative Equilibrium Operator, i.e.

$$(4.6) \quad L^2(\mathbb{P}) \ni H \quad \mapsto \quad \mathcal{S}(g_H) \stackrel{\text{def}}{=} \arg \min_{\substack{(x, \theta) \in X \\ x \in [0, C]}} \mathbb{E} \left[\left(H - x - \int_0^T \theta_t dS_t \right)^2 \right] \in X$$

by a Generative Equilibrium Operator \mathcal{G} of rank $R = 11$, depth $L = 10$, and width $M = 40$. We thus choose the non-orthogonal basis $(e_{j_1, j_2, j_3})_{j_1, j_2, j_3 \in \mathbb{N}}$ of X given by $e_{j_1, j_2, j_3}(t) \stackrel{\text{def}}{=} t^{j_1} \ln(S_t/S_0)^{j_2} (V_t/V_0)^{j_3}$, $t \in [0, T]$. Moreover, we apply the Adam algorithm over 5000 epochs with learning rate $5 \cdot 10^{-5}$ and batchsize 100 to train the Generative Equilibrium Operator on a training set consisting of 200 European call options $H_k \stackrel{\text{def}}{=} \max(S_T - K_k, 0)$, $k = 1, \dots, 200$, and 200 European put options $H_k \stackrel{\text{def}}{=} \max(K_k - S_T, 0)$, $k = 201, \dots, 400$. In addition, we evaluate its generalization performance every 125-th epoch on a test set consisting of 50 European gap call options $H_k \stackrel{\text{def}}{=} (S_T - K_{k,1}) \mathbb{1}_{\{S_T \geq K_{k,2}\}}$, $k = 401, \dots, 450$, and 50 European gap put options $H_k \stackrel{\text{def}}{=} (K_{k,2} - S_T) \mathbb{1}_{\{S_T \leq K_{k,1}\}}$, $k = 451, \dots, 500$. Hereby, the parameters $K_1, \dots, K_{400} \geq 0$ and $K_{401,1}, K_{401,2}, \dots, K_{500,1}, K_{500,2} \geq 0$ are randomly initialized with $K_{k,1} \leq K_{k,2}$, whereas the optimal prices and hedging strategies $(x_k, \theta_k) = \mathcal{S}(g_{H_k})$ are computed with the help of the minimal equivalent local martingale measure (see [62, 66]) and the Fourier arguments in [18]. The results are reported in Figure 4.3.

5. Conclusion. This paper helps close the gap between neural operator theory which suggests that solving infinite-dimensional problems requires exponentially large models, whereas reasonably sized neural operators have consistently succeeded in experimental practice. We address this gap by designing a generative neural operator, the GEO model, whose internal architecture efficiently encodes proximal forward-backward splitting algorithms at scale. Our main results (Theorems 3.2 and 3.3) demonstrate that this architecture does not suffer from the theory–practice gap: it can uniformly approximate the (approximate) solution operator for infinite families of convex splitting problems of the form (1.1), to arbitrary accuracy, with complexity that scales only logarithmically in the residual approximation error.

To illustrate the broad scope of our results, we show that solution operators for a broad class of problems—including parametric nonlinear PDEs (Section 4.1), stochastic optimal control problems (Section 4.2), and dynamic hedging problems under liquidity constraints in mathematical finance (Section 4.3), can all be cast in the form (1.1), and are therefore tractable using small GEOs. Each of these theoretical claims is also validated through empirical experiments.

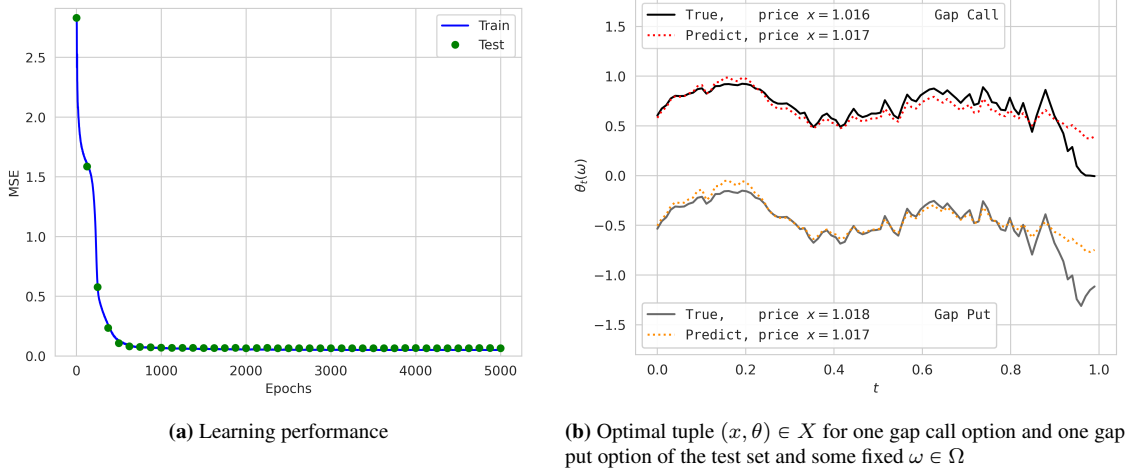


Fig. 4.3: Learning the pricing/hedging operator \mathcal{S} in (4.6) by a Generative Equilibrium Operator \mathcal{G} . In (a), the learning performance is displayed in terms of the mean squared error (MSE) $\frac{1}{|K|} \sum_{k \in K} \|\mathcal{S}(g_{H_k}) - \mathcal{G}(g_{H_k})\|^2$ on the training set (label “Train”) and test set (label “Test”). In (b), the predicted solution $\mathcal{G}(g_{H_k})$ (label “Predict”) is compared to the true solution $\mathcal{S}(g_{H_k})$ (label “True”) for one k of the test set and some $\omega \in \Omega$.

Acknowledgements and funding. A. Kratsios gratefully acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grant Nos. RGPIN-2023-04482 and DGECR-2023-00230. A. Neufeld gratefully acknowledges financial support by the MOE AcRF Tier 2 Grant MOE-2EP20222-0013. We further acknowledge that resources used in the preparation of this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and the industry sponsors of the Vector Institute³. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET) and the Digital Research Alliance of Canada (<https://alliancecan.ca/en>).

6. Proof of Theorem 3.2. We now prove our main results in a sequence of steps.

6.1. Step 0 - Idealized Forward-Backwards Splitting Scheme. We first recall and reformat the main results of [33] to our setting. Briefly, these produce a quadratically (or in some cases linearly) convergent sequence in X , converging to an optimizer of $\ell_{f,g}$, as defined in (1.1); for any suitable f and g .

LEMMA 6.1 (Convergence of the Proximal FB-Splitting Scheme with Identity Perturbations). *Suppose that $f \in \Gamma_0(X)$, $g \in C^1(X)$, and ∇g is λ -Lipschitz for some $\lambda > 1$. Fix sequences $(\lambda_l)_{l=0}^\infty$ and $(\alpha_l)_{l=0}^\infty$ in $(0, 1]$ and in $(0, 1/\lambda)$, respectively. For any $x_0 \in X$, and each $l \in \mathbb{N}_+$ define the sequence $(x_l)_{l=0}^\infty$ in X by the FB proximal splitting iteration*

$$(6.1) \quad x_{l+1} \stackrel{\text{def.}}{=} (1 - \alpha_l) x_l + \alpha_l \text{prox}_f(x_l - \lambda_l \nabla g(x_l)).$$

If for all $l \in \mathbb{N}_+$ we have $\lambda_l \leq 1/(\lambda \alpha_l)$ and if $\sup_{l \in \mathbb{N}} \|x_l\| < \infty$ then: for every time-horizon $L \in \mathbb{N}_+$

$$(6.2) \quad \ell_{f,g}(x_L) - \inf_{x \in X} \ell_{f,g}(x) \lesssim \frac{1}{L}$$

³<https://vectorinstitute.ai/partnerships/current-partners/>

where \lesssim hides an (positive) absolute constant (independent of f).

Moreover, $(x_l)_{l=0}^\infty$ converges weakly in X to a minimizer of $\ell_{f,g}$.

Before continuing with the proof of Theorem 3.2, we take a moment to establish Proposition 3.1. For every $\eta > 0$ define the operator $S_\eta : \Omega \times C^1(X) \rightarrow X$ sending any $\omega \in \Omega$ and $g \in C^1(X)$ to

$$(6.3) \quad S_\eta(\omega, g) \stackrel{\text{def.}}{=} x_L^g$$

where $x \stackrel{\text{def.}}{=} \xi(\omega)$, ξ is as in Definition 2.1, and x_L^g is as in (6.1) with $x_0 \stackrel{\text{def.}}{=} x = \xi(\omega)$. This explicitly defines the operator in Proposition 3.1; note that its initial condition is intentionally coupled to that of the Generative Equilibrium Operator (meaning that their iterations always start at the same initial condition).

Proof of Proposition 3.1. Set $L \stackrel{\text{def.}}{=} \lceil \eta \rceil$. Then, the conclusion follows from Lemma 6.1 as well as the definitions of \mathcal{X}_λ and of S_η .

To establish the continuity of $S_\eta(\omega, \cdot)$, it is enough to show the continuity of one of its iterates. Indeed, since the proximal operators prox_f is 1-Lipschitz and the map $C^1(X) \ni g \mapsto \nabla g \in C(X, X)$ is continuous with respect to the semi-norm topology τ on $C^1(X)$ (see above (2.4)), we have for every $g, \tilde{g} \in C^1(X, \mathbb{R})$ that

$$\begin{aligned} & \left\| ((1 - \alpha_l)x + \alpha_l \text{prox}_f(x - \lambda_l \nabla g(x))) - ((1 - \alpha_l)x + \alpha_l \text{prox}_f(x - \lambda_l \nabla \tilde{g}(x))) \right\|_X \\ &= \alpha_l \left\| \text{prox}_f(x - \lambda_l \nabla g(x)) - \text{prox}_f(x - \lambda_l \nabla \tilde{g}(x)) \right\|_X \\ &\leq \alpha_l \text{Lip}(\text{prox}_f) \left\| x - \lambda_l \nabla g(x) - x - \lambda_l \nabla \tilde{g}(x) \right\|_X \\ &= \alpha_l \lambda_l \left\| \nabla g(x) - \nabla \tilde{g}(x) \right\|_X \\ &\leq \max_{j=1, \dots, J} \sup_{u \in K_j} \alpha_l \lambda_l \left\| \nabla g(u) - \nabla \tilde{g}(u) \right\|_X + |g(u) - \tilde{g}(u)| \\ &= \max_{j=1, \dots, J} p_{K_j}(g - \tilde{g}) \\ &= p_{\cup_{j=1}^J K_j}(g - \tilde{g}). \end{aligned}$$

Therefore, for each $t \in \mathbb{N}_+$ and every $x \in X$, the map

$$(6.4) \quad C^1(X, \mathbb{R}) \ni g \mapsto (1 - \alpha_t)x + \alpha_t \text{prox}_f(x - \lambda_t \nabla g(x))$$

is continuous. The continuity of $S_\eta(\omega, \cdot)$ now follows. \square

Proof. For any $x_0 \in X$, define the sequence obtained by a forward-backwards (FB) proximal splitting iteration with a convex combination of the current and previous step, iteratively for each $l \in \mathbb{N}_+$ by (6.1).

Our first objective is re-expressing the FB iteration in (6.1) as in [33]. We first consider the 2-duality mapping $J_2 : X \rightarrow X^* \cong X$, defined for each $x \in X$ by

$$(6.5) \quad J_2(x) = \{x^* \in X^* \mid \langle x^*, x \rangle = \|x^*\| \|x\|, \|x^*\| = \|x\|\}.$$

As shown in [20, Proposition 4.8, page 29], it holds for every $x \in X$ that $J_2(x) = \partial(\frac{1}{2}\|\cdot\|^2)(x)$. Moreover, since we are in the Hilbert (not the general Banach) case when there is a single element in the sub-differential set $\partial(\frac{1}{2}\|\cdot\|^2)(x)$; namely the identity map id_X ; thus

$$(6.6) \quad J_2(x) = x$$

for all $x \in X$. Note that for every $f \in \Gamma_0(X)$, $g \in C^1(X)$, $\lambda > 0$, and each $x \in X$

$$\text{prox}_f(x - \lambda \nabla g(x)) = \underset{y \in X}{\text{argmin}} f(y) + \frac{1}{2} \|y - (x - \lambda \nabla g(x))\|^2$$

$$\begin{aligned}
&= \operatorname{argmin}_{y \in X} f(y) + \frac{1}{2} \|y\|^2 - \frac{2}{2} \langle y, x - \lambda \nabla g(x) \rangle \\
&= \operatorname{argmin}_{y \in X} f(y) + \frac{1}{2} \|y\|^2 - \langle y, x \rangle + \langle y, \lambda \nabla g(x) \rangle \\
(6.7) \quad &= \operatorname{argmin}_{y \in X} \frac{1}{2} \|y - x\|^2 + \lambda \langle y, \nabla g(x) + J_2(0) \rangle + f(y).
\end{aligned}$$

Consequently, (6.1) and (6.7) imply that for each $l \in \mathbb{N}_+$ we have

$$(6.8) \quad x_{l+1} \stackrel{\text{def}}{=} (1 - \alpha_l) x_l + \alpha_l \left(\operatorname{argmin}_{y \in X} \frac{1}{2} \|y - x_l\|^2 + \lambda \langle y, \nabla g(x_l) + J_2(z_l) \rangle + f(y) \right)$$

where $z_l \stackrel{\text{def}}{=} 0$ for all $l \in \mathbb{N}_+$. Now, under the boundedness assumption $\sup_{l \in \mathbb{N}} \|x_l\| < \infty$ and since $0 < \lambda_l < 1/\lambda$ then [33, Proposition 2 (iii)] implies that (6.2) holds and [33, Proposition 2 (iii)] guarantees that $(x_l)_{l=0}^\infty$ converges weakly to a minimizer of $\ell_{f,g}$ in X . \square

6.2. Step 1 - Approximately Implementing the Gradient Operator. Since, in general, we cannot assume that we can directly implement the gradient operator ∇ , our first step is to approximate it via a finite difference as follows. Recall that the Gâteaux derivative Df of a real-valued function f on X which is Gâteaux differentiable function at some $x \in X$ is given by

$$(6.9) \quad Df(x)(y) \stackrel{\text{def}}{=} \lim_{\eta \downarrow 0} \frac{f(x + \eta y) - f(x)}{\eta}.$$

We denote by $C^1(X)$ the set of functions $f : X \rightarrow \mathbb{R}$ which are Gâteaux differentiable functions at all points in X with bounded Gâteaux derivative. If f is Gâteaux differentiable at x , then its *Fréchet gradient* $\nabla f(x)$, see e.g. [7, Remark 2.55], must satisfy

$$(6.10) \quad Df(x)(y) = \langle y, \nabla f(x) \rangle$$

for each $y \in X$. Upon fixing an orthonormal basis $(e_i)_{i \in I}$ of X , we may re-write the right-hand side of (6.10) by

$$(6.11) \quad Df(x)(y) = \left\langle \sum_{i \in I} \langle y, e_i \rangle e_i, \nabla f(x) \right\rangle = \sum_{i \in I} \langle y, e_i \rangle \langle e_i, \nabla f(x) \rangle.$$

By definition of both derivatives, we have: for each $i \in I$

$$(6.12) \quad \langle e_i, \nabla f(x) \rangle = Df(x)(e_i) = \lim_{\eta \downarrow 0} \frac{f(x + \eta e_i) - f(x)}{\eta} = (\partial_t f(x + t e_i))|_{t=0}.$$

Consequently, (6.11) implies that $Df(x)(y) = \sum_{i \in I} \langle y, e_i \rangle (\partial_t f(x + t e_i))|_{t=0}$; whence

$$(6.13) \quad \nabla f(x) = \sum_{i \in I} (\partial_t f(x + t e_i))|_{t=0} e_i.$$

Without loss of generality, we identify I with an initial segment of \mathbb{N} . We consider the case where I is infinite, with the case where $\#I < \infty$ being more straightforward but similar; whence, for us $I = \mathbb{N}$.

This motivates our finite-rank, finite-difference operator. Fix a rank $R \in \mathbb{N}_+$ and a precision parameter $\delta > 0$. Given any $f \in C^1(X)$. We approximate the Fréchet gradient in (6.13) by the *rank R - δ -divided difference operator* Δ_δ^R . In what follows, we consider the δ -divided difference operator defined for each $f \in C^1(X)$ at every $x \in X$ by

$$(6.14) \quad \Delta_\delta^R(f)(x) \stackrel{\text{def}}{=} \sum_{i=0}^{R-1} \frac{f(x + \delta e_i) - f(x)}{\delta} e_i.$$

As one may expect, the rank R - δ -divided difference operator provides a rank R approximation to the Fréchet gradient of any $C^1(X)$ function. Interestingly, if the target function's gradient's "higher frequencies" (coefficients of the e_i for large i) are exponentially small, then R need only grow logarithmically in the reciprocal approximation error $\varepsilon > 0$.

LEMMA 6.2 (Finite Difference Approximation of Gradient Operator). *Suppose that $f \in C^1(X)$ and let \mathcal{K} be a non-empty compact subset of X . If $\delta > 0$ and $R \in \mathbb{N}$, then*

$$(6.15) \quad \sup_{x \in \mathcal{K}} \|\Delta_\delta^R(f)(x) - \nabla f(x)\|_X \lesssim R\delta + \sqrt{\sum_{i=R}^{\infty} |(\partial_t f(x + te_i))|_{t=0}|^2}.$$

For instance, if there exist constants $r, C > 0$ such that: for all $x \in \mathcal{K}$ and $i \in \mathbb{N}$ we have $|(\partial_t f(x + te_i))|_{t=0}|^2 \leq C e^{-2r i}$, then there is a constant $c > 0$ such that for every $\varepsilon > 0$ we may pick δ small enough and $R \in \mathcal{O}(c + \log(1/\varepsilon))$ satisfying

$$(6.16) \quad \sup_{x \in \mathcal{K}} \|\Delta_\delta^R(f) - \nabla f(x)\|_X \leq \varepsilon.$$

Proof of Lemma 6.2. Suppose that: there are $C, r > 0$ such that for all $x \in \mathcal{K}$ and each $i \in \mathbb{N}_+$

$$(6.17) \quad |\langle \nabla f(x), e_i \rangle| \leq C e^{-ri}.$$

Fix $\delta > 0$. Then, for every $x \in \mathcal{K}$, we have

$$(6.18) \quad \begin{aligned} \|\Delta_\delta^R(f)(x) - \nabla f(x)\|_X &\leq \left\| \sum_{i=0}^{R-1} \left(\frac{f(x + \delta e_i) - f(x)}{\delta} - (\partial_t f(x + te_i))|_{t=0} \right) e_i \right. \\ &\quad \left. + \sum_{i=R}^{\infty} (\partial_t f(x + te_i))|_{t=0} e_i \right\|_X \\ &\leq \sum_{i=0}^{R-1} \left\| \frac{f(x + \delta e_i) - f(x)}{\delta} - (\partial_t f(x + te_i))|_{t=0} \right\| \|e_i\|_X \\ &\quad + \sqrt{\sum_{i=R}^{\infty} ((\partial_t f(x + te_i))|_{t=0})^2 \|e_i\|_X^2} \\ &\leq R\tilde{C}\delta + \sqrt{\sum_{i=R}^{\infty} |(\partial_t f(x + te_i))|_{t=0}|^2}, \end{aligned}$$

where we have used Taylor's theorem/standard 1-dimensional finite (forward) difference estimates to obtain (6.18); where $0 \leq \tilde{C} \stackrel{\text{def}}{=} \sup_{(x,t) \in X \times [0,\delta]} |(\partial_t f(x + te_i))|$ and $\tilde{C} < \infty$ by the continuity of ∇f and the compactness of $\mathcal{K} \times [0, \delta]$. Now, if there are constant $C, r > 0$ such that $|(\partial_t f(x + te_i))|_{t=0}|^2 \leq C e^{-2r i}$ for all $i \in \mathbb{N}$; then

$$(6.19) \quad \sqrt{\sum_{i=R}^{\infty} |(\partial_t f(x + te_i))|_{t=0}|^2} \leq \sqrt{C} \sqrt{\sum_{i=R}^{\infty} e^{-i2r}} = \frac{\sqrt{C}}{\sqrt{1 - e^{2r}}} e^{-Rr}.$$

Setting $C' \stackrel{\text{def}}{=} \max\{\sqrt{C}/\sqrt{1 - e^{2r}}, \tilde{C}\}$ yields the bound

$$(6.20) \quad \sup_{x \in X} \|\Delta_\delta^R(f)(x) - \nabla f(x)\|_X \leq C' R\delta + C' e^{-Rr}.$$

Let $\varepsilon > 0$ be given. Retroactively setting $R \stackrel{\text{def}}{=} \lceil \ln((2C')^{1/r}) + \frac{1}{r} \ln(\varepsilon^{-1}) \rceil$ and $\delta = \varepsilon/(2C'R)$ completes our proof. \square

6.3. Step 2 - Approximate Implementation of Proximal Forward-Backwards Splitting. We first approximate the idealized proximal forward-backward splitting scheme considered in Lemma 6.1 by a variant where the Fréchet gradient operator is replaced by a finite-rank finite-difference variance. Thus, the next lemma takes the “differentiation” component of our problem one step closer to an implementable object on a computer processing finite-dimensional linear algebra.

LEMMA 6.3 (Approximate Proximal Splitting Scheme). *Let $R \in \mathbb{N}_+$, and $\delta, \lambda > 0$, $f \in C^1(X)$ and suppose that ∇g is λ -Lipschitz. Let $(\alpha_l)_{l=0}^\infty, (\lambda_l)_{l=0}^\infty$ be as in Lemma 6.1. Define the approximate proximal splitting iteration, for each $l \in \mathbb{N}_+$ by*

$$(6.21) \quad \hat{x}_{l+1} \stackrel{\text{def}}{=} (1 - \alpha_l)\hat{x}_l + \alpha_l \text{prox}_f(\hat{x}_l - \lambda_l \Delta_\delta^R(g)(\hat{x}_l)).$$

Then, for every $L \in \mathbb{N}_+$ we have

$$(6.22) \quad \|x_L - \hat{x}_L\|_X \lesssim (R\delta + \tau(R, g))(1 - 2^{-(L+1)})$$

where $\tau(R, g)^2 \stackrel{\text{def}}{=} \sum_{i=R}^\infty |(\partial_t g(x + te_i))|_{t=0}|^2$.

Proof of Lemma 6.3. Recall that prox_f is firmly non-expansive, see e.g. [7, Proposition 12.28]; thus it is 1-Lipschitz. Consequently, for every $l \in \mathbb{N}_+$ we have

$$\begin{aligned} & \|x_{l+1} - \hat{x}_{l+1}\|_X \\ & \leq \|(1 - \alpha_l)x_l + \alpha_l \text{prox}_f(x_l - \lambda_l \nabla g(x_l)) - (1 - \alpha_l)\hat{x}_l - \alpha_l \text{prox}_f(\hat{x}_l - \lambda_l \Delta_\delta^R(g)(\hat{x}_l))\|_X \\ & \leq (1 - \alpha_l)\|x_l - \hat{x}_l\|_X + \alpha_l \|\text{prox}_f(x_l - \lambda_l \nabla g(x_l)) - \text{prox}_f(\hat{x}_l - \lambda_l \Delta_\delta^R(g)(\hat{x}_l))\|_X \\ & \leq (1 - \alpha_l)\|x_l - \hat{x}_l\|_X + \alpha_l \|(x_l - \lambda_l \nabla g(x_l)) - (\hat{x}_l - \lambda_l \Delta_\delta^R(g)(\hat{x}_l))\|_X \\ & \leq (1 - \alpha_l)\|x_l - \hat{x}_l\|_X + \alpha_l \|x_l - \hat{x}_l\|_X + \alpha_l \lambda_l \|\nabla g(x_l) - \Delta_\delta^R(g)(\hat{x}_l)\|_X \\ & \leq (1 - \alpha_l)\|x_l - \hat{x}_l\|_X + \alpha_l \|x_l - \hat{x}_l\|_X + \alpha_l \lambda_l \|\nabla g(x_l) - \nabla g(\hat{x}_l)\|_X + \alpha_l \lambda_l \|\nabla g(\hat{x}_l) - \Delta_\delta^R(g)(\hat{x}_l)\|_X \\ & \leq (1 - \alpha_l)\|x_l - \hat{x}_l\|_X + \alpha_l \|x_l - \hat{x}_l\|_X + \alpha_l \lambda_l \|x_l - \hat{x}_l\|_X + \alpha_l \lambda_l \|\nabla g(\hat{x}_l) - \Delta_\delta^R(g)(\hat{x}_l)\|_X \\ & = (1 + \alpha_l \lambda_l)\|x_l - \hat{x}_l\|_X + \alpha_l \lambda_l \|\nabla g(\hat{x}_l) - \Delta_\delta^R(g)(\hat{x}_l)\|_X \end{aligned} \tag{6.23}$$

$$\leq (1 + \frac{1}{\lambda})\|x_l - \hat{x}_l\|_X + \alpha_l \lambda_l \|\nabla g(\hat{x}_l) - \Delta_\delta^R(g)(\hat{x}_l)\|_X$$

$$(6.24) \quad \leq 2\|x_l - \hat{x}_l\|_X + \underbrace{\alpha_l \lambda_l \|\nabla g(\hat{x}_l) - \Delta_\delta^R(g)(\hat{x}_l)\|_X}_{(I)},$$

where (6.23) held by assumption that for each $l \in \mathbb{N}_+$ we had $\alpha_l \in [0, 1]$ and that $\lambda_l \in (0, 1/\lambda)$. Now, under our assumptions, term (I) can be bounded using Lemma 6.2. Let $\tau(R, g)^2 \stackrel{\text{def}}{=} \sum_{i=R}^\infty |(\partial_t g(x + te_i))|_{t=0}|^2$. The right-hand side of (6.24) can further be controlled, for every $l \in \mathbb{N}_+$, by

$$(6.25) \quad \|x_{l+1} - \hat{x}_{l+1}\|_X \leq 2\|x_l - \hat{x}_l\|_X + \alpha_l \lambda_l (R\delta + \tau(R, g)).$$

Recursively applying the estimate in (6.25) we find that

$$\|x_L - \hat{x}_L\|_X \leq 2^L \|x_0 - \hat{x}_0\|_X + (R\delta + \tau(R, g)) \sum_{s=0}^L \alpha_s \lambda_s \prod_{u=s}^t (1 + \alpha_u \lambda_u \lambda_g)$$

$$(6.26) \quad \leq 2^L \underbrace{\|x_0 - \hat{x}_0\|_X}_{(\text{II})} + (R\delta + \tau(R, g)) \sum_{s=0}^L \alpha_s \lambda_s 2^{(L-s-1)+}.$$

Note that $x_0 = \hat{x}_0$; whence, (II) vanishes. Next, if for each $l \in \mathbb{N}_+$ with $l \leq L$, we constrain $\alpha_l \leq \lambda_l / 2^{L-2l}$ then (6.26) can be further controlled as

$$\begin{aligned} \|x_L - \hat{x}_L\|_X &\leq (R\delta + \tau(R, g)) \sum_{s=0}^L \alpha_s \lambda_s 2^{(L-s-1)+} \\ &\leq (R\delta + \tau(R, g)) \sum_{s=0}^L \frac{1}{2^{L-2s}} \lambda_s 2^{(L-s-1)+} \\ &\leq (R\delta + \tau(R, g)) \sum_{s=0}^L \frac{1}{2^s} \\ &\leq 2(R\delta + \tau(R, g))(1 - 2^{-(L+1)}). \end{aligned} \quad \square$$

Unfortunately, the proximal operator prox_f need not map $\text{span}(\{e_j\}_{j=0}^{R-1})$ into itself. Thus, we further modify the iteration in Lemma 6.3 to incorporate a projection step following the application of prox_f back down onto the span of $\text{span}(\{e_j\}_{j=0}^{R-1})$.

LEMMA 6.4 (Approximation by Projected (Finite-Dimensional) Proximal Splitting Scheme). *Let $R \in \mathbb{N}_+$, $\delta, \lambda > 0$, $f \in C^1(X)$ and suppose that ∇g is λ -Lipschitz. Let $(\alpha_l)_{l=0}^\infty, (\lambda_l)_{l=0}^\infty$ be as in Lemma 6.1. Define the approximate proximal splitting iteration for each $l \in \mathbb{N}_+$ by*

$$(6.27) \quad z_{l+1} \stackrel{\text{def.}}{=} (1 - \alpha_l)z_l + \alpha_l \sigma_f(z_l - \lambda_l \Delta_\delta^R(g)(z_l)).$$

Then, for every $L \in \mathbb{N}_+$, if the hyperparameters $\alpha_0, \dots, \alpha_T$ satisfy

$$(6.28) \quad 0 < \alpha_l \leq 2^{-l-L} \left(\max \left\{ \sup_{u \in \text{prox}_f(K)} \|\text{prox}_f(u) - P_R(u)\|_X, 1 \right\} \right)^{-(L-l-1)+}$$

then, for each $l = 0, \dots, L$ we have

$$(6.29) \quad \|\hat{x}_L - z_L\| \leq 2^{1-L},$$

where $C_r > 0$ depends only on r .

The proof of Lemma 6.4 relies on the following two technical lemmata elucidating some elementary properties of the operator Δ_δ^R .

LEMMA 6.5 (Finite Difference-Type Operator Δ_δ^R are Bounded). *Let $\delta > 0$, $\lambda \geq 0$, $R \in \mathbb{N}_+$, and $g : \mathcal{X} \rightarrow \mathbb{R}$ be λ -Lipschitz. Then, $\Delta_\delta^R(g)$ is $\frac{2R\lambda}{\delta}$ -Lipschitz.*

Proof. Let $x, \tilde{x} \in X$. Then,

$$\begin{aligned} \|\Delta_\delta^R(g)(x) - \Delta_\delta^R(g)(\tilde{x})\|_X &= \frac{1}{\delta} \left\| \sum_{i=0}^{R-1} g(x + \delta e_i) - g(x) - g(\tilde{x} + \delta e_i) + g(\tilde{x}) \right\| \\ &\leq \frac{1}{\delta} \sum_{i=0}^{R-1} (|g(x + \delta e_i) - g(\tilde{x} + \delta e_i)| + |g(x) - g(\tilde{x})|) \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\lambda}{\delta} \sum_{i=0}^{R-1} (\|x + \delta e_i - \tilde{x} + \delta e_i\|_X + \|x - \tilde{x}\|_X) \\
&= \frac{2\lambda R}{\delta} \|x - \tilde{x}\|_X.
\end{aligned}$$

Thus, Δ_δ^R is $(2\lambda R)/\delta$ -Lipschitz. \square

LEMMA 6.6 (Projection Operator Approximation Properties). *Let $r > 0$, $C \geq 0$, $R \in \mathbb{N}_+$ and let $\mathcal{K} \stackrel{\text{def.}}{=} \{x \in X : |\langle x, e_i \rangle| \leq C e^{-ri}\}$. Then the projection operator $P_R : X \mapsto \text{span}\{e_i\}_{i=0}^{R-1}$ satisfies*

- (i) $P_R(\mathcal{K}) \subseteq \mathcal{K}$ and
 - (ii) $\sup_{x \in \mathcal{K}} \|P_R(x) - x\|_X \leq C_r e^{-rR}$,
- where $C_r \stackrel{\text{def.}}{=} C/\sqrt{1 - e^{2r}} > 0$.

Proof. Let $x \in \mathcal{K}$. Then by linearity of P_R ,

$$P_R(x) = \sum_{i=0}^{\infty} \langle x, e_i \rangle P_R(e_i) = \sum_{i=0}^{\infty} \langle x, e_i \rangle e_i I_{i < R} = \sum_{i=0}^{R-1} \langle x, e_i \rangle e_i.$$

Therefore, for each $i \in \mathbb{N}$, we have $|\langle P_R(x), e_i \rangle| \leq C e^{-ri} I_{i < R} \leq C e^{-ri}$. Thus, $P_R(x) \in \mathcal{K}$ and (i) is verified. Moreover,

$$\|P_R(x) - x\| = \sqrt{\sum_{i=0}^{R-1} |\langle x, e_i \rangle|^2 \|e_i\|^2} \leq C \sqrt{\sum_{i=0}^{R-1} e^{-ri2}} = C_r e^{-Rr}.$$

Thus, (ii) holds. \square

Proof of Lemma 6.4. Fix $R \in \mathbb{N}_+$ and $\delta > 0$. Then, for every $t \in \mathbb{N}_+$, we have that

$$\begin{aligned}
(6.30) \quad &\|\hat{x}_l - z_l\|_X \\
&\leq (1 - \alpha_l) \|x_{l-1} - z_{l-1}\|_X \\
&+ \alpha_l \|\text{prox}_f(\hat{x}_{l-1} - \lambda_l \Delta_\delta^R(g)(\hat{x}_{l-1})) - P_R \circ \text{prox}_f(z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1}))\|_X \\
&\leq (1 - \alpha_l) \|x_{l-1} - z_{l-1}\|_X \\
&+ \alpha_l \|\text{prox}_f(\hat{x}_{l-1} - \lambda_l \Delta_\delta^R(g)(\hat{x}_{l-1})) - \text{prox}_f(z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1}))\|_X \\
&+ \alpha_l \|\text{prox}_f(z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1})) - P_R \circ \text{prox}_f(z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1}))\|_X \\
&\leq (1 - \alpha_l) \|x_{l-1} - z_{l-1}\|_X \\
&+ \alpha_l \text{Lip}(\text{prox}_f) \|\hat{x}_{l-1} - \lambda_l \Delta_\delta^R(g)(\hat{x}_{l-1}) - z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1})\|_X \\
&+ \alpha_l \|\text{prox}_f(z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1})) - P_R \circ \text{prox}_f(z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1}))\|_X \\
&\leq (1 - \alpha_l) \|x_{l-1} - z_{l-1}\|_X \\
&+ \alpha_l \text{Lip}(\text{prox}_f) \|\hat{x}_{l-1} - \lambda_l \Delta_\delta^R(g)(\hat{x}_{l-1}) - z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1})\|_X
\end{aligned}$$

$$(6.31) \quad + \alpha_l \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X$$

$$\begin{aligned}
(6.32) \quad &\leq (1 - \alpha_l) \|x_{l-1} - z_{l-1}\|_X \\
&+ \alpha_l \|\hat{x}_{l-1} - \lambda_l \Delta_\delta^R(g)(\hat{x}_{l-1}) - z_{l-1} - \lambda_l \Delta_\delta^R(g)(z_{l-1})\|_X \\
&+ \alpha_l \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X
\end{aligned}$$

$$\begin{aligned}
&\leq (1 - \alpha_l) \|x_{l-1} - z_{l-1}\|_X \\
&+ \alpha_l \|\lambda_l \Delta_\delta^R(g)(\hat{x}_{l-1}) - \lambda_l \Delta_\delta^R(g)(z_{l-1})\|_X + \alpha_l \|\hat{x}_{l-1} - z_{l-1}\|_X \\
&+ \alpha_l \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X \\
&= \|x_{l-1} - z_{l-1}\|_X + \alpha_l \lambda_l \|\Delta_\delta^R(g)(\hat{x}_{l-1}) - \Delta_\delta^R(g)(z_{l-1})\|_X \\
&+ \alpha_l \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X \\
(6.33) \quad &\leq \|x_{l-1} - z_{l-1}\|_X + \alpha_l \lambda_l \frac{\alpha_l 2R}{\delta} \|\hat{x}_{l-1} - z_{l-1}\|_X + \alpha_l \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X \\
(6.34) \quad &\leq \left(1 + \frac{2R}{\delta}\right) \|x_{l-1} - z_{l-1}\|_X + \alpha_l \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X,
\end{aligned}$$

where we used (6.32) again held by the firm non-expansiveness of prox_f (see, e.g., [7, Proposition 12.28]), implying that prox_f is 1-Lipschitz, we used Lemma 6.5 to deduce (6.33), and we used the constant $\alpha_l \leq 1$ to deduce (6.34).

Moreover, by compactness of \mathcal{K} and by continuity of prox_f , we have that $\text{prox}_f(\mathcal{K})$ is compact. Thus, by the metric approximation property in separable Hilbert spaces we have that

$$(6.35) \quad C_{1:R} \stackrel{\text{def.}}{=} \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X < \infty.$$

Fix a time-horizon $L \in \mathbb{N}_+$. Incorporating (6.35) in the right-hand side of (6.34) and iterating we obtain the bound

$$(6.36) \quad \|\hat{x}_L - z_L\|_X \leq \underbrace{\prod_{l=0}^L \left(1 + \frac{2R}{\delta}\right)}_{\text{(III)}} \|x_0 - z_0\|_X + \underbrace{\sum_{l=0}^L \alpha_s C_{1:R}^{(L-l-1)+}}_{\text{(IV)}}.$$

Now, since $x_0 \in E_R$ we may indeed pick $x_0 = z_0$; implying that (III) vanishes. Likewise, if

$$(6.37) \quad \alpha_l C_{1:R}^{(L-l-1)+} \leq \frac{1}{2^{L+l}}$$

for each $l = 0, \dots, L$ then $\sum_{s=0}^l \alpha_l C_{1:R}^{(l-s-1)+} \leq \frac{1}{2^L} \sum_{s=0}^l \frac{1}{2^s} \leq 2^{1-L}$. Now the constraint (6.37) is equivalent to the condition (6.28). Consequently, (IV) is also controllable and the estimate (6.36) reduces to

$$\|\hat{x}_L - z_L\|_X \leq \frac{1}{2^L} \sum_{s=0}^L \frac{1}{2^s} \leq 2^{1-L}. \quad \square$$

6.4. Step 4 - Convergence Under Additional Regularity of f and g . Under more regularity on f and on the input g , we may guarantee that the neural operator is minimizing the loss function.

PROPOSITION 6.7 (Convergence of Objective Function). *Fix $x \in X$. Let $\lambda, \lambda_f, \lambda_g, \delta > 0$, $L, R \in \mathbb{N}_+$, $f \in \Gamma(X)$ be coercive and bounded from below. Let $(\alpha_l)_{l=0}^\infty, (\lambda_l)_{l=0}^\infty$ be sequences satisfying the conditions of Lemma 6.1 and the decay condition*

$$(6.38) \quad 0 < \alpha_l \leq 2^{-l-L} \left(\max \left\{ \sup_{u \in \text{prox}_f(\mathcal{K})} \|\text{prox}_f(u) - P_R(u)\|_X, 1 \right\} \right)^{-(L-l-1)+}$$

and let $(x_l)_{l=0}^\infty$ and $(z_l)_{l=0}^\infty$ be given by (6.21) and (6.27), respectively, with $x_0 \stackrel{\text{def.}}{=} z_0 \stackrel{\text{def.}}{=} x \in X$. Then, for any $g \in C^1(X)$ with λ -Lipschitz Fréchet gradient

$$(6.39) \quad \|x_l - z_l\|_X \lesssim 2^{1-L} + (R\delta + \tau(R, g))(1 - 2^{-(L+1)})$$

and \lesssim hides a constant independent of δ, R, g , and L . If, additionally, f is λ_f -Lipschitz and if g is λ_g -Lipschitz with λ -Lipschitz Fréchet gradient then (6.39) strengthens to

$$(6.40) \quad \ell_{f,g}(z_T) - \inf_{x \in X} \ell_{f,g}(x) \lesssim \frac{1}{L} + (\lambda_f + \lambda_g) \left(2^{1-L} + (R\delta + \tau(R, g))(1 - 2^{-(L+1)}) \right)$$

where $\tau(R, g)^2 \stackrel{\text{def.}}{=} \sum_{i=R}^\infty |(\partial_t g(x + te_i))|_{t=0}|^2$.

Proof of Proposition 6.7. We first establish (6.39); indeed

$$(6.41) \quad \begin{aligned} &\leq \|z_L - x_L\|_X \\ &\leq \|z_L - \hat{x}_L\|_X + \|x_L - \hat{x}_L\|_X \end{aligned}$$

$$(6.42) \quad \leq 2^{1-L} + \|x_L - \hat{x}_L\|_X$$

$$(6.43) \quad \lesssim 2^{1-L} + (R\delta + \tau(R, g))(1 - 2^{-(L+1)}),$$

where (6.42) held by Lemma 6.4 due to our decay assumptions on α , made in (6.29), and (6.43) held by Lemma (6.3) by our assumptions on the Lipschitzness of the gradient of g .

Next, we establish (6.40). We first show the existence of a minimizer to $\ell_{f,g}$ over X , which we will routinely use momentarily. Since f was assumed to be coercive, then for every $\eta \in \mathbb{R}$ the level set $f^{-1}[(-\infty, \eta]]$ is relatively compact in X (see e.g. [23, Definition 1.12]). Since g was assumed to take non-negative values then, the level set $(f + g)^{-1}[(-\infty, \eta]] \subseteq f^{-1}[(-\infty, \eta]]$ is relatively compact; i.e. $f + g$ is coercive. Now, since f is also bounded below, then Tonelli's direct method, see e.g. [23, Theorem 1.15], implies that there exists a minimizer $x_{f,g}^* \in X$ of $\ell_{f,g}$; i.e.

$$(6.44) \quad \inf_{x \in X} \ell_{f,g}(x) = \ell_{f,g}(x_{f,g}^*).$$

Now, using the Lipschitzness of $f + g$ and the minimality of $x_{f,g}^*$ in (6.44) we have

$$(6.45) \quad \begin{aligned} \ell_{f,g}(z_L) - \inf_{x \in X} \ell_{f,g}(x) &= |\ell_{f,g}(z_L) - \inf_{x \in X} \ell_{f,g}(x)| \\ &= |\ell_{f,g}(z_L) - \ell_{f,g}(x_{f,g}^*)| \\ &\leq |\ell_{f,g}(z_L) - \ell_{f,g}(x_L)| + |\ell_{f,g}(x_L) - \ell_{f,g}(x_{f,g}^*)| \\ &= |\ell_{f,g}(z_L) - \ell_{f,g}(x_L)| + \ell_{f,g}(x_L) - \ell_{f,g}(x_{f,g}^*) \\ &\lesssim |\ell_{f,g}(z_L) - \ell_{f,g}(x_L)| + \frac{1}{L} \end{aligned}$$

$$(6.46) \quad \lesssim (\lambda_f + \lambda_g) \left(2^{1-L} + (R\delta + \tau(R, g))(1 - 2^{-(L+1)}) \right) + \frac{1}{L}$$

where (6.45) held by Lemma 6.1, (6.41) held by our Lipschitzness assumptions on f and on g , and (6.46) held by (6.39). \square

We are now in place to establish Theorem 3.2. Indeed, we only need to show that z_L (as defined in (6.27)) can be computed by a Generative Equilibrium Operator of depth L and we only need to verify that x_L (as defined in (6.1)) is the output of $S_\eta(\omega, g)$ (as defined in (6.3)). We prove both our main theorems together, as this yields the most streamlined treatment thereof.

Proofs of Theorem 3.2 and 3.3. Fix $\omega \in \Omega$, $\eta, > 0$, and let $S_\eta \stackrel{\text{def.}}{=} S_\eta(\omega, \cdot) : C^1(X) \rightarrow X$ be defined as in (6.3). Set $x_0 \stackrel{\text{def.}}{=} z_0 \stackrel{\text{def.}}{=} \xi(\omega) \in X$ and couple

$$\delta \stackrel{\text{def.}}{=} 2^{-L}/R > 0.$$

Fix any $R \in \mathbb{N}_+$, set $M \stackrel{\text{def.}}{=} R$, $L \stackrel{\text{def.}}{=} \lceil 1/\eta \rceil$, and for each $l \in \{1, \dots, L\}$ fix any $\alpha_l, \lambda_l \in (0, 1/\lambda)$ such that $(\alpha_l)_{l=1}^L$ satisfies the decay condition in (6.38). For each $l \in \{1, \dots, L\}$ we iteratively define the GEO layers $\mathcal{L}^{(l)}$ (see Definition 2.1) by

$$(6.47) \quad \mathcal{L}_g^{(l)}(x) \stackrel{\text{def.}}{=} \gamma^{(l)}x + (1 - \gamma^{(l)})\sigma_f\left(A^{(l)}x + [B^{(l)}(g(x + x_m^{(l)}))_{m=1}^M + b^{(l)}]^{\uparrow:M}\right)$$

where, for each $m = 1, \dots, R$, we set $x_m^{(l)} \stackrel{\text{def.}}{=} e_m$, $A^{(l)} = I_R$ (the $R \times R$ identity matrix) and $B^{(l)} \stackrel{\text{def.}}{=} \frac{\lambda_l}{\delta} I_D$, $b^{(l)} = \mathbf{0}_R$ (the zero vector in \mathbb{R}^R), and $\gamma_l \stackrel{\text{def.}}{=} 1 - \alpha_l$. Then, by definition of rank R , δ -divided difference operator $\Delta_\delta^R(\cdot)$ (defined in (6.14)), the lifting/embedding operator $\cdot^{\uparrow:M}$ (defined in (2.2)), and each GEO layer $\mathcal{L}^{(l)}$ in (6.47) we have that

$$(6.48) \quad \mathcal{L}_g^{(l)}(x) \stackrel{\text{def.}}{=} (1 - \alpha^{(l)})x + \gamma^{(l)}\sigma_f(x + \lambda_l \Delta_\delta^R(x)).$$

Consequently, we find that

$$(6.49) \quad \mathcal{L}_g^{(L)} \circ \dots \circ \mathcal{L}_g^{(1)}(x_0) = z_L$$

where z_L is defined in (6.21). Now, observe that $\mathcal{G}(\omega, g) \stackrel{\text{def.}}{=} \mathcal{L}^{(L)} \circ \dots \circ \mathcal{L}^{(1)}(x_0)$ is a well-defined GEO (with dependence on ω implicitly in $x_0 = \xi(\omega)$ and on g by the definitions of each GEO layer in (6.47)). Consequently, Proposition 6.7 and the definition of the $\mathcal{O}(\eta)$ -approximate solution operator S_η in (6.3) imply that

$$(6.50) \quad \sup_{g \in \mathcal{X}_\lambda} \|S_\eta(\omega, g) - \mathcal{G}(\omega, g)\|_X \lesssim 2^{1-L} + (2^{-L} + \tau(R, g))(1 - 2^{-(L+1)})$$

with \lesssim hiding a constant independent of δ , R , L (and thus of η), and of any $g \in \mathcal{X}_\lambda$. Restricting the supremum in (6.39) to the set $\mathcal{X}_\lambda(r)$ (defined in (3.2)) we find that

$$(6.51) \quad \sup_{g \in \mathcal{X}_\lambda(r)} \|S_\eta(\omega, g) - \mathcal{G}(\omega, g)\|_X \lesssim 2^{1-L} + (2^{-R} + r 2^{-R})(1 - 2^{-(L+1)})$$

$$(6.52) \quad \begin{aligned} &\leq 2^{1-L} + (2^{-R} + r 2^{-R}) \\ &\lesssim 2^{-L} + 2^{-R}. \end{aligned}$$

Fix an approximation error $\varepsilon > 0$. Retroactively, setting $R \stackrel{\text{def.}}{=} L \stackrel{\text{def.}}{=} \lceil \varepsilon \rceil$; then (6.51)-(6.52) implies that

$$(6.53) \quad \sup_{g \in \mathcal{X}_\lambda(r)} \|S_\eta(\omega, g) - \mathcal{G}(\omega, g)\|_X \lesssim \varepsilon$$

yielding (3.3). If, additionally, f is λ_f -Lipschitz and if g is λ_g -Lipschitz with λ -Lipschitz Fréchet gradient then (6.40) in Lemma 6.7 yields (3.4). \square

References.

- [1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Applied mathematics series / National Bureau of Standards 55, Print. 9, Dover, New York, 9th ed., 1970.
- [2] B. ADCOCK, N. DEXTER, AND S. MORAGA SCHEUERMANN, *Optimal deep learning of holomorphic operators between Banach spaces*, Advances in Neural Information Processing Systems, 37 (2024), pp. 27725–27789.
- [3] G. ALVAREZ, I. EKREN, A. KRATSIOS, AND X. YANG, *Neural operators can play dynamic Stackelberg games*, arXiv preprint arXiv:2411.09644, (2024).
- [4] K. AZIZZADENESHELI, N. KOVACHKI, Z. LI, M. LIU-SCHIAFFINI, J. KOSSAIFI, AND A. ANANDKUMAR, *Neural operators for accelerating scientific simulations and design*, Nature Reviews Physics, 6 (2024), pp. 320–328.
- [5] A. BACHOUCH, C. HURÉ, N. LANGRENÉ, AND H. PHAM, *Deep neural networks algorithms for stochastic control problems on finite horizon: Numerical applications*, Methodology and Computing in Applied Probability, 24 (2022), pp. 143–178.
- [6] S. BAI, J. Z. KOLTER, AND V. KOLTUN, *Deep equilibrium models*, in Advances in Neural Information Processing Systems, vol. 32, 2019.
- [7] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, Cham, second ed., 2017, <https://doi.org/10.1007/978-3-319-48311-5>. With a foreword by Hedy Attouch.
- [8] C. BECK, S. BECKER, P. CHERIDITO, A. JENTZEN, AND A. NEUFELD, *Deep splitting method for parabolic PDEs*, SIAM Journal on Scientific Computing, 43 (2021), pp. A3135–A3154.
- [9] S. BECKER, P. CHERIDITO, AND A. JENTZEN, *Deep optimal stopping*, Journal of Machine Learning Research, 20 (2019), pp. 1–25.
- [10] J. A. L. BENITEZ, T. FURUYA, F. FAUCHER, A. KRATSIOS, X. TRICOCHÉ, AND M. V. DE HOOP, *Out-of-distributional risk bounds for neural operators with applications to the Helmholtz equation*, Journal of Computational Physics, 513 (2024), p. 113168.
- [11] Y. BENYAMINI AND J. LINDENSTRAUSS, *Geometric nonlinear functional analysis. Vol. 1*, vol. 48 of American Mathematical Society Colloquium Publications, American Mathematical Society, Providence, RI, 2000, <https://doi.org/10.1090/coll/048>.
- [12] J. BERNER, M. DABLANDER, AND P. GROHS, *Numerically solving parametric families of high-dimensional Kolmogorov partial differential equations via deep learning*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 16615–16627.
- [13] D. P. BERTSEKAS, *Dynamic programming and optimal control*, Athena scientific series in optimization and computation, Athena Scientific, Belmont, Mass, third ed., 2005.
- [14] K. BHATTACHARYA, B. HOSSEINI, N. B. KOVACHKI, AND A. M. STUART, *Model reduction and neural networks for parametric PDEs*, The SMAI Journal of computational mathematics, 7 (2021), pp. 121–157.
- [15] H. S. D. O. BORDE, A. LUKOIANOV, A. KRATSIOS, M. BRONSTEIN, AND X. DONG, *Scalable message passing neural networks: No need for attention in large graph representation learning*, arXiv preprint arXiv:2411.00835, (2024).
- [16] K. BREDIES, *A forward-backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space*, Inverse Problems, 25 (2009), pp. 015005, 20, <https://doi.org/10.1088/0266-5611/25/1/015005>.
- [17] H. BUEHLER, L. GONON, J. TEICHMANN, AND B. W. AND, *Deep hedging*, Quantitative Finance, 19 (2019), pp. 1271–1291.
- [18] P. CARR AND D. B. MADAN, *Option valuation using the fast Fourier transform*, Journal of Computational Finance, 2 (1999), pp. 61–73.

- [19] Y. CHEN, Y. SHI, AND B. ZHANG, *Optimal control via neural networks: A convex approach*, arXiv preprint arXiv:1805.11835, (2018).
- [20] I. CIORANESCU, *Geometry of Banach spaces, duality mappings and nonlinear problems*, vol. 62 of Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1990, <https://doi.org/10.1007/978-94-009-2121-4>.
- [21] P. L. COMBETTES AND V. R. WAJS, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200, <https://doi.org/10.1137/050626090>.
- [22] C. CUCHIERO, P. SCHMOCKER, AND J. TEICHMANN, *Global universal approximation of functional input maps on weighted spaces*, arXiv preprint 2306.03303, (2023).
- [23] G. DAL MASO, *An introduction to Γ -convergence*, vol. 8 of Progress in Nonlinear Differential Equations and their Applications, Birkhäuser Boston, Inc., Boston, MA, 1993, <https://doi.org/10.1007/978-1-4612-0327-8>.
- [24] W. E, M. HUTZENTHALER, A. JENTZEN, AND T. KRUSE, *On multilevel Picard numerical approximations for high-dimensional nonlinear parabolic partial differential equations and high-dimensional nonlinear backward stochastic differential equations*, Journal of Scientific Computing, 79 (2019), p. 1534=1571.
- [25] W. E, M. HUTZENTHALER, A. JENTZEN, AND T. KRUSE, *Multilevel picard iterations for solving smooth semi-linear parabolic heat equations*, Partial Differential Equations and Applications, 2 (2021), p. 80.
- [26] M. FEISCHL, C. SCHWAB, AND F. ZEHETGRUBER, *Neural general operator networks via Banach fixed point iterations*, tech. report, ETH Zurich, Research Report No. 2025-13, 2025.
- [27] W. H. FLEMING AND R. W. RISHEL, *Deterministic and stochastic optimal control*, Stochastic Modelling and Applied Probability, 1, Springer-Verlag, Berlin, Germany, 1st ed., 1975.
- [28] T. FURUYA AND A. KRATSIOS, *Simultaneously solving FBSDEs with neural operators of logarithmic depth, constant width, and sub-linear rank*, arXiv preprint arXiv:2410.14788, (2024).
- [29] M. GABOR, T. PIOTROWSKI, AND R. L. G. CAVALCANTE, *Positive concave deep equilibrium models*, arXiv preprint arXiv:2402.04029, (2024).
- [30] L. GALIMBERTI, A. KRATSIOS, AND G. LIVIERI, *Designing universal causal deep learning models: The case of infinite-dimensional dynamical systems from stochastic analysis*, arXiv preprint arXiv:2210.13300, (2022).
- [31] M. GEIST, P. PETERSEN, M. RASLAN, R. SCHNEIDER, AND G. KUTYNIOK, *Numerical solution of the parametric diffusion equation by deep neural networks*, Journal of Scientific Computing, 88 (2021), p. 22.
- [32] A. GNOATTO, S. LAVAGNINI, AND A. PICARELLI, *Deep quadratic hedging*, Mathematics of Operations Research, 0 (0), p. null.
- [33] W.-B. GUAN AND W. SONG, *The forward–backward splitting method and its convergence rate for the minimization of the sum of two functions in Banach spaces*, Optimization Letters, 15 (2021), pp. 1735–1758.
- [34] J. HAN AND W. E, *Deep learning approximation for stochastic control problems*, arXiv preprint arXiv:1611.07422, (2016).
- [35] J. HAN, A. JENTZEN, AND W. E, *Solving high-dimensional partial differential equations using deep learning*, Proceedings of the National Academy of Sciences, 115 (2018), pp. 8505–8510.
- [36] M. HERDE, B. RAONIC, T. ROHNER, R. KÄPPELI, R. MOLINARO, E. DE BÉZENAC, AND S. MISHRA, *Poseidon: Efficient foundation models for PDEs*, Advances in Neural Information Processing Systems, 37 (2024), pp. 72525–72624.
- [37] I. KARATZAS AND S. E. SHREVE, *Methods of mathematical finance*, Probability theory and stochastic modelling, volume 39, Springer, New York, corrected 4th printing 2016 ed., 2016.
- [38] Y. KHOO, J. LU, AND L. YING, *Solving parametric PDE problems with artificial neural networks*, European Journal of Applied Mathematics, 32 (2021), p. 421–435.
- [39] Y. KOROLEV, *Two-layer neural networks with values in a Banach space*, SIAM Journal on Mathematical Analysis, 54 (2022), pp. 6358–6389.

- [40] N. KOVACHKI, S. LANTHALER, AND S. MISHRA, *On universal approximation and error bounds for Fourier neural operators*, Journal of Machine Learning Research, 22 (2021), pp. 1–76.
- [41] N. KOVACHKI, Z. LI, B. LIU, K. AZIZZADENESHELI, K. BHATTACHARYA, A. STUART, AND A. ANANDKUMAR, *Neural operator: Learning maps between function spaces with applications to PDEs*, Journal of Machine Learning Research, 24 (2023), pp. 1–97.
- [42] A. KRATSIOS, T. FURUYA, J. A. L. BENITEZ, M. LASSAS, AND M. DE HOOP, *Mixture of experts soften the curse of dimensionality in operator learning*, arXiv preprint arXiv:2404.09101, (2024).
- [43] G. KUTYNIOK, P. PETERSEN, M. RASLAN, AND R. SCHNEIDER, *A theoretical analysis of deep neural networks and parametric PDEs*, Constructive Approximation, 55 (202), pp. 73–125.
- [44] S. LANTHALER, *Operator learning with PCA-Net: upper and lower complexity bounds*, Journal of Machine Learning Research, 24 (2023), pp. 1–67.
- [45] S. LANTHALER, *Operator learning of Lipschitz operators: An information-theoretic perspective*, arXiv preprint arXiv:2406.18794, (2024).
- [46] S. LANTHALER, S. MISHRA, AND G. E. KARNIADAKIS, *Error estimates for DeepONets: A deep learning framework in infinite dimensions*, Transactions of Mathematics and Its Applications, 6 (2022), p. tnac001.
- [47] X. LI, D. VERMA, AND L. RUTHOTTO, *A neural network approach for stochastic optimal control*, SIAM Journal on Scientific Computing, 46 (2024), pp. C535–C556.
- [48] Z. LI, N. KOVACHKI, K. AZIZZADENESHELI, B. LIU, K. BHATTACHARYA, A. STUART, AND A. ANANDKUMAR, *Fourier neural operator for parametric partial differential equations*, arXiv preprint arXiv:2010.08895, (2020).
- [49] A. E. B. LIM, *Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market*, Mathematics of Operations Research, 29 (2004), pp. 132–161.
- [50] L. LU, P. JIN, G. PANG, Z. ZHANG, AND G. E. KARNIADAKIS, *Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators*, Nature Machine Intelligence, 3 (2021), pp. 218–229.
- [51] C. MARCATI AND C. SCHWAB, *Exponential convergence of deep operator networks for elliptic partial differential equations*, SIAM Journal on Numerical Analysis, 61 (2023), pp. 1513–1545.
- [52] T. MARWAH, A. POKLE, J. Z. KOLTER, Z. C. LIPTON, J. LU, AND A. RISTESKI, *Deep equilibrium based neural operators for steady-state PDEs*, in Thirty-seventh Conference on Neural Information Processing Systems, 2023, <https://openreview.net/forum?id=v6YzxwJlQn>.
- [53] S. MOHAMMAD-TAHERI, M. J. COLBROOK, AND S. BRUGIAPAGLIA, *Deep greedy unfolding: Sorting out argsorting in greedy sparse recovery algorithms*, arXiv preprint arXiv:2505.15661, (2025).
- [54] R. MOLINARO, Y. YANG, B. ENGQUIST, AND S. MISHRA, *Neural inverse operators for solving PDE inverse problems*, in Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 25105–25139.
- [55] V. MONGA, Y. LI, AND Y. C. ELDAR, *Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing*, IEEE Signal Processing Magazine, 38 (2021), pp. 18–44.
- [56] N. MUÇA CIRONE AND C. SALVI, *Rough kernel hedging*, arXiv preprint arXiv:2501.09683, (2025).
- [57] A. NEUFELD AND P. SCHMOCKER, *Chaotic hedging with iterated integrals and neural networks*, arXiv preprint arXiv:2209.10166, (2022).
- [58] A. NEUFELD AND P. SCHMOCKER, *Universal approximation property of Banach space-valued random feature models including random neural networks*, arXiv preprint arXiv:2312.08410, (2023).
- [59] A. NEUFELD, P. SCHMOCKER, AND S. WU, *Full error analysis of the random deep splitting method for nonlinear parabolic PDEs and PIDEs*, Communications in Nonlinear Science and Numerical Simulation, 143 (2025), p. 108556.
- [60] L. NEYT, J. TOFT, AND J. VINDAS, *Hermite expansions for spaces of functions with nearly optimal time-frequency decay*, J. Funct. Anal., 288 (2025), pp. Paper No. 110706, 17, <https://doi.org/10.1016/j.jfa.2024.110706>.

- [61] J. PATHAK, S. SUBRAMANIAN, L. BERKELEY, P. HARRINGTON, S. RAJA, M. MARDANI, T. KURTH, D. HALL, Z. LI, K. AZIZZADENESHELI, ET AL., *Fourcastnet: A global data-driven high-resolution weather model using adaptive Fourier neural operators*, Ann Arbor, 1001 (2022), p. 48109.
- [62] H. PHAM, *On quadratic hedging in continuous time*, Mathematical Methods of Operations Research, 51 (2000), pp. 315–339.
- [63] M. A. RAHMAN, M. A. FLOREZ, A. ANANDKUMAR, Z. E. ROSS, AND K. AZIZZADENESHELI, *Generative adversarial neural operators*, arXiv preprint arXiv:2205.03017, (2022).
- [64] R. H. RIEDI, R. BALESTRIERO, AND R. G. BARANIUK, *Singular value perturbation and deep network optimization*, Constructive Approximation, 57 (2023), pp. 807–852.
- [65] J. RUF AND W. WANG, *Hedging with linear regressions and neural networks*, Journal of Business & Economic Statistics, 40 (2022), pp. 1442–1454.
- [66] M. SCHWEIZER, *A guided tour through quadratic hedging approaches*, SFB 373 Discussion Papers 1999,96, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, 1999.
- [67] H. M. SONER, *Stochastic optimal control in finance*, Cattedra Galileiana, Scuola Normale Superiore, Pisa, 2005.
- [68] N. TOUZI, *Optimal Stochastic Control, Stochastic Target Problems, and Backward SDE*, Springer, New York, NY, 2014.
- [69] E. WINSTON AND J. Z. KOLTER, *Monotone operator equilibrium networks*, Advances in neural information processing systems, 33 (2020), pp. 10718–10728.
- [70] Y. YANG, A. F. GAO, J. C. CASTELLANOS, Z. E. ROSS, K. AZIZZADENESHELI, AND R. W. CLAYTON, *Seismic wave propagation and inversion with neural operators*, The Seismic Record, 1 (2021), pp. 126–134.

Appendix A. Supplementary Material.

A.1. Examples of Proximal Operators. For some prominent Hilbert spaces X and functions $f : X \rightarrow (-\infty, \infty]$, we compute the corresponding activation function $\sigma_f : X \rightarrow X$ defined as $\sigma_f(x) \stackrel{\text{def.}}{=} P_R(\text{prox}_f(x))$, for $x \in X$, where $R \in \mathbb{N}_+$. For example, on any Hilbert space X , the proximal operator of $f(x) \stackrel{\text{def.}}{=} \frac{1}{2}\|x\|^2$ is given by $\text{prox}_f(x) = \frac{1}{2}x$. Hence, we obtain a linear R -rank operator $\sigma_f(x) = P_R(\text{prox}_f(x)) = \frac{1}{2}P_R(x)$ as activation function.

EXAMPLE A.1. For $d \in \mathbb{N}_+$, let $X \stackrel{\text{def.}}{=} \mathbb{R}^d$, set $R \stackrel{\text{def.}}{=} d - 1$, and define $f : X \rightarrow (-\infty, \infty]$ by $f(x) = 0$ if $x \in [0, \infty)^d$, and $f(x) = \infty$ otherwise. Then, for every $x \stackrel{\text{def.}}{=} (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, it holds that $\text{prox}_f(x) = \text{ReLU}_d(x) \stackrel{\text{def.}}{=} (\max(x_1, 0), \dots, \max(x_d, 0))^\top$. Hence, by using that $P_R = \text{id}_{\mathbb{R}^d}$, we obtain the multivariate ReLU activation function

$$\sigma_f(x) = \text{prox}_f(x) = \text{ReLU}_d(x).$$

Moreover, the linear operators $A^{(l)} \in L(\mathbb{R}^d; \mathbb{R}^d) \cong \mathbb{R}^{d \times d}$ and $B^{(l)} \in \mathbb{R}^{d \times M}$ in Definition 2.1 correspond to matrices, while $b^{(l)} \in \mathbb{R}^d$ are classical bias vectors.

EXAMPLE A.2. For the sequence space $X \stackrel{\text{def.}}{=} l^2 \stackrel{\text{def.}}{=} \{x \stackrel{\text{def.}}{=} (x_i)_{i \in \mathbb{N}} : \|x\| \stackrel{\text{def.}}{=} \sum_{i=0}^{\infty} x_i^2 < \infty\}$ and some fixed $R \in \mathbb{N}_+$, we define $f : X \rightarrow (-\infty, \infty]$ by $f(x) = \|x\|_{l^1} \stackrel{\text{def.}}{=} \sum_{i=1}^{\infty} |x_i|$, for $x \in l^2$. Then, for every $x \stackrel{\text{def.}}{=} (x_i)_{i \in \mathbb{N}} \in l^2$, it holds that $\text{prox}_f(x) = ((x_i + 1)\mathbb{1}_{\{x_i < -1\}} + (x_i - 1)\mathbb{1}_{\{x_i > 1\}})_{i \in \mathbb{N}}$. Thus, by using the projection P_R , we obtain a non-linear R -rank activation function

$$\begin{aligned} \sigma_f(x) &= \sum_{j=0}^{R-1} \langle \text{prox}_f(x), e_j \rangle e_j \\ &= ((x_0 + 1)\mathbb{1}_{\{x_0 < -1\}} + (x_0 - 1)\mathbb{1}_{\{x_0 > 1\}}, \dots, (x_{R-1} + 1)\mathbb{1}_{\{x_{R-1} < -1\}} + (x_{R-1} - 1)\mathbb{1}_{\{x_{R-1} > 1\}}, 0, 0, \dots). \end{aligned}$$

Moreover, the linear operators $A^{(l)} \in L(l_R^2; l_R^2) \cong \mathbb{R}^{R \times R}$ in Definition 2.1 are of the form $l_R^2 \ni x \stackrel{\text{def.}}{=} (x_0, \dots, x_{R-1}, 0, 0, \dots) \mapsto A^{(l)}x = (\sum_{j=0}^{R-1} a_{0,j}^{(l)} x_j, \dots, \sum_{j=0}^{R-1} a_{R-1,j}^{(l)} x_j, 0, 0, \dots) \in l_R^2$ for some $a^{(l)} \stackrel{\text{def.}}{=} (a_{i,j}^{(l)})_{i,j=0,\dots,R-1} \in \mathbb{R}^{R \times R}$, while $B^{(l)} \in \mathbb{R}^{R \times M}$ and $b^{(l)} \in \mathbb{R}^R$ are classical matrices and vectors.

EXAMPLE A.3. For the L^2 -space $X \stackrel{\text{def.}}{=} L^2(\mu) \stackrel{\text{def.}}{=} L^2(\Omega, \mathcal{A}, \mu)$, a basis $(e_j)_{j \in \mathbb{N}}$ of $L^2(\mu)$, some fixed $R \in \mathbb{N}$, and some $-\infty < c_1 < c_2 < \infty$, we define the function $f : L^2(\mu) \rightarrow (-\infty, \infty]$ by $f(x) = 0$ if $x(\Omega) \subseteq [c_1, c_2]$, and $f(x) \stackrel{\text{def.}}{=} \infty$ otherwise. Then, the proximal operator $\text{prox}_f(x) = \text{proj}_{[c_1, c_2]}(x(\cdot))$ is the pointwise projection to $[c_1, c_2]$ defined by $\text{proj}_{[c_1, c_2]}(u) \stackrel{\text{def.}}{=} \min(\max(u, c_1), c_2)$ for $u \in \mathbb{R}$. Hence, we obtain a non-linear R -rank activation function

$$\sigma_f(x) = \sum_{j=0}^{R-1} \langle \text{prox}_f(x), e_j \rangle e_j = \sum_{j=0}^{R-1} \langle \text{proj}_{[c_1, c_2]}(x(\cdot)), e_j \rangle e_j.$$

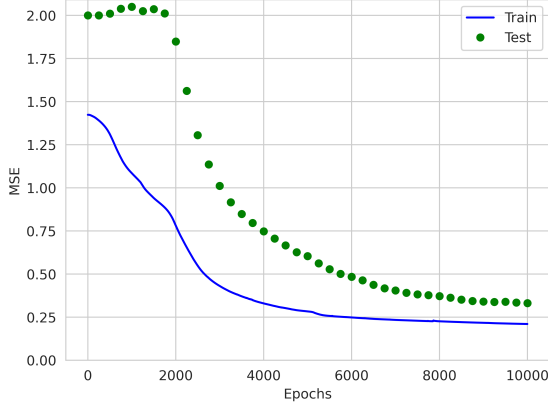
Moreover, the linear operators $A^{(l)} \in L(L^2(\mu)_R; L^2(\mu)_R) \cong \mathbb{R}^{R \times R}$ in Definition 2.1 are of the form $L^2(\mu)_R \ni x \mapsto A^{(l)}x = \sum_{i,j=0}^{R-1} a_{i,j}^{(l)} \langle x, e_i \rangle e_j \in L^2(\mu)_R$ for some $a^{(l)} \stackrel{\text{def.}}{=} (a_{i,j}^{(l)})_{i,j=0,\dots,R-1} \in \mathbb{R}^{R \times R}$, while $B^{(l)} \in \mathbb{R}^{R \times M}$ and $b^{(l)} \in \mathbb{R}^R$ are classical matrices and vectors.

A.2. A Finite-Dimensional Application: Learning a minimization operator. As an additional sanity check, we first consider a splitting problem over a finite dimensional Hilbert space. For the Hilbert space

$X = \mathbb{R}^d$ and a convex subset $C \subseteq \mathbb{R}^d$, we aim to learn the operator

$$\{g : \mathbb{R}^d \rightarrow \mathbb{R} \text{ is convex}\} \ni g \mapsto \arg \min_{x \in C} g(x) = \arg \min_{x \in \mathbb{R}^d} (f(x) + g(x)) \in \mathbb{R}^d,$$

where $f : \mathbb{R}^d \rightarrow (-\infty, \infty]$ is defined by $f(x) = 0$ if $x \in C$, and $f(x) = \infty$ otherwise. In this case, the proximal operator of f is given as $\text{prox}_f(x) = \text{proj}_C(x) \stackrel{\text{def.}}{=} \arg \min_{y \in C} \|x - y\|$, for all $x \in \mathbb{R}^d$.



(a) Learning performance

k	True	Predict
9001	$\begin{pmatrix} 1.000 \\ 1.000 \end{pmatrix}$	$\begin{pmatrix} 0.998 \\ 0.985 \end{pmatrix}$
9002	$\begin{pmatrix} 1.000 \\ 1.000 \end{pmatrix}$	$\begin{pmatrix} 0.889 \\ 0.916 \end{pmatrix}$
9501	$\begin{pmatrix} -1.000 \\ -1.000 \end{pmatrix}$	$\begin{pmatrix} -0.692 \\ -1.000 \end{pmatrix}$
9502	$\begin{pmatrix} -1.000 \\ -1.000 \end{pmatrix}$	$\begin{pmatrix} -1.000 \\ -0.961 \end{pmatrix}$

(b) Solution of (A.1) for four functions g_k of the test set.

Fig. A.1: Learning the minimization operator \mathcal{S} in (A.1) by a Generative Equilibrium Operator \mathcal{G} . In (a), the learning performance is displayed in terms of the mean squared error (MSE) $\frac{1}{|K|} \sum_{k \in K} \|\mathcal{S}(g_k) - \mathcal{G}(g_k)\|^2$ on the training set (label “Train”) and test set (label “Test”). In (b), the predicted solution $\mathcal{G}(g_k)$ (label “Predict”) is compared to the true solution $\mathcal{S}(g_k)$ (label “True”) for four k of the test set.

EXAMPLE A.4. For $d = 2$, we consider the Hilbert space $X = \mathbb{R}^d$ and the convex subset $C = [-1, 1]^d$. In this setting, we aim to learn the minimization operator

$$(A.1) \quad \{g : \mathbb{R}^d \rightarrow \mathbb{R} \text{ is convex}\} \ni g \mapsto \mathcal{S}(g) \stackrel{\text{def.}}{=} \arg \min_{x \in [-1, 1]^d} g(x) = \arg \min_{x \in \mathbb{R}^d} (f(x) + g(x)) \in \mathbb{R}^d,$$

by a Generative Equilibrium Operator \mathcal{G} of rank $R = d = 2$, depth $L = 20$, and sample points $M = 20$. Hereby, the function $f : \mathbb{R}^d \rightarrow (-\infty, \infty]$ is defined as above.

To this end, we choose the standard orthonormal basis of \mathbb{R}^d . Moreover, we apply the Adam algorithm over 10000 epochs with learning rate $2 \cdot 10^{-4}$ to train the Generative Equilibrium Operator on a training set consisting of 9000 convex functions $\mathbb{R}^d \ni x \mapsto g_k(x) \stackrel{\text{def.}}{=} \frac{1}{2} x^\top A_k x + b_k^\top x + c_k \in (-\infty, \infty)$, $k = 1, \dots, 9000$, where $A_k \in \mathbb{S}_+^d$, $b_k \in \mathbb{R}^d$, and $c_k \in \mathbb{R}$ are randomly initialized. In addition, we evaluate its generalization performance every 250-th on a test set consisting of 1000 convex functions $\mathbb{R}^d \ni x \mapsto g_k(x) \stackrel{\text{def.}}{=} \ln \left(\sum_{i=1}^d \exp(b_{k,i} x_i) + c_k \right) \in (-\infty, \infty)$, $k = 9001, \dots, 10000$, where $b_k \stackrel{\text{def.}}{=} (b_{k,1}, \dots, b_{k,d})^\top \in \mathbb{R}^d$ (with either $b_{k,i} \geq 0$ for all $i = 1, \dots, d$, or $b_{k,i} \leq 0$ for all $i = 1, \dots, d$) and $c_k \in \mathbb{R}$ are also randomly initialized. The results are reported in Figure A.1.