PDE-constrained Models with Neural Network Terms: Optimization and Global Convergence

Justin Sirignano,[*] Jonathan MacArt,[†] and Konstantinos Spiliopoulos[‡][§]

March 29, 2022

Abstract

Recent research has used deep learning to develop partial differential equation (PDE) models in science and engineering. The functional form of the PDE is determined by a neural network, and the neural network parameters are calibrated to available data. Calibration of the embedded neural network can be performed by optimizing over the PDE. Motivated by these applications, we rigorously study the optimization of a class of linear elliptic PDEs with neural network terms. The neural network parameters in the PDE are optimized using gradient descent, where the gradient is evaluated using an adjoint PDE. As the number of parameters become large, the PDE and adjoint PDE converge to a non-local PDE system. Using this limit PDE system, we are able to prove convergence of the neural network-PDE to a global minimum during the optimization. Finally, we use this adjoint method to train a neural network model for an application in fluid mechanics, in which the neural network functions as a closure model for the Reynolds-averaged Navier–Stokes (RANS) equations. The RANS neural network model is trained on several datasets for turbulent channel flow and is evaluated out-of-sample at different Reynolds numbers.

1 Introduction

Consider the optimization problem where we seek to minimize the objective function

$$J(\theta) = \frac{1}{2} \sum_{\ell=1}^{L} \left(\langle u, m_\ell \rangle - \langle h, m_\ell \rangle \right)^2, \tag{1.1}$$

where h is a target profile, $\{m_\ell\}_{\ell=1}^{L}$ are given functions that could be interpreted as moments or eigenfunctions for example. Here, we denote $\langle u, m_\ell \rangle = \int_U u(x) m_\ell(x) dx$. The variable u satisfies the elliptic partial differential equation (PDE)

$$\begin{aligned} Au(x) &= f(x, \theta), \qquad x \in U \\ u(x) &= 0, \qquad x \in \partial U, \end{aligned} \tag{1.2}$$

where A is a standard second-order elliptic operator and $U \subset \mathbb{R}^d$. Thus, we wish to select the parameters θ such that the solution u of the PDE (1.2) is as close as possible to the target profile h.

If $A^\dagger$ denotes the formal adjoint operator to A, then the adjoint PDE is

$$\begin{aligned} A^\dagger \hat{u}(x) &= (\langle u, m \rangle - \langle h, m \rangle)^\top m(x), \qquad x \in U \\ \hat{u}(x) &= 0, \qquad x \in \partial U, \end{aligned} \tag{1.3}$$

where $m(x) = (m_1(x), m_2(x), \ldots, m_L(x))$ and $(\langle u, m \rangle - \langle h, m \rangle)^\top m(x) = \sum_{\ell=1}^{L} (\langle u, m_\ell \rangle - \langle h, m_\ell \rangle) m_\ell(x)$.

[*]Mathematical Institute at the University of Oxford, E-mail: Justin.Sirignano@maths.ox.ac.uk

[†]Department of Aerospace & Mechanical Engineering at the University of Notre Dame, E-mail: jmacart@nd.edu

[‡]Department of Mathematics & Statistics at Boston University, Corresponding Author, E-mail: kspiliop@math.bu.edu

The optimization of (1.1) with the PDE constraint (1.2) using the adjoint equation (1.3) is a classic PDE-constrained optimization problem. Adjoint methods have been widely used for the optimization of PDE models for engineering systems: see, for example, [2–4, 8, 15, 17, 19–22, 28–33, 37, 41, 50–52, 54, 59, 67, 68]. [5] recently numerically optimized a linear PDE with a neural network source term using adjoint methods. [49, 62], [26], [66], and [42] use adjoint methods to calibrate nonlinear PDE models with neural network terms. [5, 26, 42, 49, 62, 66] do not include a mathematical analysis of the adjoint method for PDE models with neural network terms.

We set the source term f in equation (1.2) to be a neural network f^N with N hidden units:

$$f^N(x, \theta) = \frac{1}{N^\beta} \sum_{i=1}^{N} c^i \sigma(w^i x + \eta^i), \tag{1.4}$$

where $(c^i, w^i, \eta^i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$, w^i is a column vector, x is a row vector, $\theta = \{(c^i, w^i, \eta^i), i = 1, \cdots, N\}$, and $\beta \in (\frac{1}{2}, 1)$. The factor $\frac{1}{N^\beta}$ is a normalization.

As the number of hidden units increase in the neural network (1.4), the number of parameters also increases. The neural network can represent more complex functions with larger numbers of hidden units and parameters.

The parameters θ are learned via continuous-time gradient descent:

$$\dot{\theta}_t^i = -\alpha^N \nabla_{\theta^i} J(\theta_t), \tag{1.5}$$

where α^N is the learning rate and $\theta_t^i = (c_t^i, w_t^i, \eta_t^i)$ for $i = 1, \cdots, N$. The gradient $\nabla_{\theta^i} J(\theta_t)$ is evaluated by solving the adjoint PDE (1.3) and using the formula

$$\nabla_\theta J(\theta) = \int_U \hat{u}(x) \nabla_\theta f(x, \theta) dx. \tag{1.6}$$

We will analyze the solution to equations (1.2), (1.3), and (1.5) when the source term is a neural network f^N and as the number of hidden units $N \to \infty$.

Define

$$g_t^N(x) = f^N(x, \theta_t) = \frac{1}{N^\beta} \sum_{i=1}^{N} c_t^i \sigma(w_t^i x + \eta_t^i). \tag{1.7}$$

Let $u(x; f)$ and $\hat{u}(x; f)$ respectively be the solutions $u(x)$ and $\hat{u}(x)$ with a forcing function f. Define $(u^N(t, x), \hat{u}^N(t, x)) = (u(x; g_t^N), \hat{u}(x; g_t^N))$. Let us also define the notation $u^N(t) = u^N(t, \cdot)$ and $\hat{u}^N(t) = \hat{u}^N(t, \cdot)$. We explicitly denote the dependence of the objective function (1.1) on the number of hidden units N in the neural network by introducing the notation

$$J_t^N \equiv J(\theta_t) = \frac{1}{2} \left\| \langle u^N - h, m \rangle \right\|_2^2, \tag{1.8}$$

where $\|\cdot\|_2$ is the Euclidean norm.

We prove that

$$\lim_{t \to \infty, N \to \infty} J_t^N = 0$$

and that the limits $t \to \infty$ and $N \to \infty$ are interchangeable.

The proof consists of three steps. First, we prove that, as $N \to \infty$, $u^N \to u^*$ and $\hat{u}^N \to \hat{u}^*$, where u^* and $\hat{u}^*$ satisfy a non-local PDE system. We then prove that the limit PDE system reaches a the global minimum; i.e., for any $\epsilon > 0$, there exists a time τ such that $\|(u^*, m)(\tau) - (m, h)\|_2 < \epsilon$. This analysis is challenging due to the lack of a spectral gap for the eigenvalues of the non-local linear operator in the limit PDE system. The third and final step is to use the analysis of the limit PDE system to prove global convergence for the pre-limit PDE system (1.2), which requires proving that the pre-limit solution remains in a neighborhood of the limit solution on some time interval $[T_0, \infty)$.

1.1 Applications in Science and Engineering

The governing equations for a number of scientific and engineering applications are well-understood. However, numerically solving these equations can be computationally intractable for many real-world applications. The exact physics can involve complex phenomena occurring over disparate ranges of spatial and temporal scales, which can require infeasibly large computational grids to accurately resolve. Examples abound in air and space flight, power generation, and biomedicine [12, 36, 63, 65, 69]. For instance, fully resolving turbulent flows in fluid mechanics can quickly become infeasible at flight-relevant Reynolds numbers.

The exact physics equations must therefore be approximated in practice, resulting in reduced-order PDE models which can be numerically solved at tractable cost. This frequently introduces substantial approximation error, and model predictions can be unreliable and fail to match observed real-world outcomes. These reduced-order PDE models are derived from the "exact" physics but introduce unclosed terms which must be approximated with a model.

Examples of reduced-order PDE models in fluid mechanics include the Reynolds-averaged Navier–Stokes (RANS) equations and Large-eddy Simulation (LES) equations. Recent research has explored using deep learning models for the closure of these equations [26, 42, 47–49, 57, 60–62, 66, 70]. In parallel, other classes of machine learning methods have also been developed for PDEs such as physics-informed neural networks (PINNs) [10, 11, 34, 38, 55, 56, 71]. An overview of the current progress in developing machine learning models in fluid mechanics can be found in [6, 7, 14].

This recent research on combining PDE models with neural networks motivates the analysis in our paper. In Section 3, we implement the adjoint method for training a neural network closure model for the RANS equations for turbulent channel flow. This numerical section demonstrates that the theoretical results of the paper are applicable beyond the linear case and it also demonstrates how the approach can be applied in practice to nonlinear differential equations as well as within the framework of a strong formulation of the objective function (1.1). The extension of our theoretical results to the nonlinear PDE case is left for future work.

1.2 Organization of the Article

Section 2 presents the main theorems regarding the global convergence of the PDE solution $u^N(t, x)$ to a global minimizer of the objective function (1.1). The example of neural network PDE models in fluid mechanics is presented in Section 3. The remaining sections contain the proof of the theorems. Section 4 contains several useful a priori bounds. The convergence $u^N \to u^*$ and $\hat{u}^N \to \hat{u}^*$ as $N \to \infty$ is proven in Section 6. The uniqueness and existence of a solution to the non-local linear PDE system is proven in Section 5. Section 7 studies the spectral properties of the limit PDE system and proves that $u^*(t)$ reaches a the global minimum; i.e., for any $\epsilon > 0$, there exists a time τ such that $\|(u^*, m)(\tau) - (m, h)\|_2 < \epsilon$. Section 8 proves the global convergence of the pre-limit PDE solution $\lim\limits_{t \to \infty, N \to \infty} J_t^N = 0$. In Section 9 we discuss possible extensions of this work.

2 Main Results

Let U be an open, bounded subset of $\mathbb{R}^d$. We will denote by $(\cdot, \cdot)$ the usual inner product in $H = L^2(U)$. We shall assume that the operator A is uniformly elliptic, diagonalizable and dissipative, per Assumption 2.1.

Assumption 2.1. The operator A is a second order uniformly elliptic operator with uniformly bounded coefficients and self-adjoint. The countable complete orthonormal basis $\{e_n\}_{n \in \mathbb{N}} \subset H$ that consists of eigenvectors of A corresponds to a non-negative sequence $\{\lambda_n\}_{n \in \mathbb{N}}$ of eigenvalues which satisfy

$$(-A)e_n = -\lambda_n e_n, \quad n \in \mathbb{N}$$

and the dissipativity condition $\lambda = \inf\limits_{n \in \mathbb{N}} \lambda_n > 0$ holds.

An example of a second-order elliptic operator is

$$Au(x) = -\sum_{i,j=1}^{n} \left(a^{i,j}(x)u_{x_j}(x) + b^i(x)u(x)\right)_{x_i} + \sum_{i=1}^{n} b^i(x)u_{x_i}(x) - c(x)u(x).$$

3

The assumption above that A is uniformly elliptic and self-adjoint guarantees that A is diagonalizable [18, Theorem 8.8.37].

Assumption 2.2. We assume that the target profile $h(x)$ and the functions $m(x)$ satisfy the following conditions:

- The target profile h satisfies $h \in L^2(U)$ and, in addition, $Ah \in L^2(U)$.

- $\{m_\ell\}_{\ell=1}^L$ are orthonormal functions and $Am_\ell \in L^2(U)$.

Notice that if $L = 1$, then we can consider any function $m(x)$ as long as $Am \in L^2(U)$ since one can always normalize by $\|m\|_{L_2}$.

Assumption 2.3. The following assumptions for the neural network (1.4) and the learning rate are imposed:

- The activation function $\sigma \in C_b^2(\mathbb{R})$, i.e. σ is twice continuously differentiable and bounded. We also assume that σ is non-constant.

- The randomly initialized parameters (C_0^i, W_0^i, η_0^i) are i.i.d., mean-zero random variables with a distribution $\mu_0(dc, dw, d\eta)$.

- The random variable C_0^i has compact support and $\langle \|w\| + |\eta|, \mu_0 \rangle = \int_{\mathbb{R}^{d+2}} (\|w\| + |\eta|) \mu_0(dc, dw, d\eta) < \infty$.

- The learning rate is chosen to be $\alpha^N = \frac{\alpha}{N^{1-2\beta}}$ where $0 < \alpha < \infty$ and $\beta \in (1/2, 1)$.

We first study the limiting behavior of the PDE system (1.2)-(1.3) when the source term f in (1.2) is $g_t^N(x)$. The pre-limit system of PDEs can be written as follows

$$
\begin{aligned}
Au^N(t, x) &= g_t^N(x), \quad x \in U \\
u^N(t, x) &= 0, \quad x \in \partial U,
\end{aligned}
\tag{2.1}
$$

and

$$
\begin{aligned}
A^\dagger \hat{u}^N(t, x) &= (\langle u^N, m \rangle - \langle h, m \rangle)^\top m(x), \quad x \in U \\
\hat{u}^N(t, x) &= 0, \quad x \in \partial U.
\end{aligned}
\tag{2.2}
$$

θ_t satisfies the differential equation

$$
\dot{\theta}_t^i = -\alpha^N \nabla_{\theta_t^i} J(\theta_t),
\tag{2.3}
$$

where $\theta_0^i \sim \mu_0$ are sampled independently for each $i = 1, \cdots, N$ and α^N is the learning rate. We will denote by Θ the space in which the initialization $\theta_0^i = (c_0^i, w_0^i, \eta_0^i)$ take values in. The gradient is evaluated via the formula (see Lemma 4.1)

$$
\nabla_{\theta_i} J(\theta_t) = \int_U \hat{u}^N(t, x) \nabla_{\theta_i} f^N(x, \theta_t) dx.
\tag{2.4}
$$

Before proceeding to describe the limiting PDE system, a few remarks on the pre-limit system (2.1)-(2.2) are in order.

Remark 2.4. Based on our assumptions, the Lax-Milgram theorem (see for example Chapter 5.8 of [18]) says that the elliptic boundary-value problem (2.1)-(2.2) has a unique weak solution $(u^N, \hat{u}^N) \in H_0^1(U) \times H_0^1(U)$. Let us also recall here that classical elliptic regularity theory guarantees that if for given $m \in \mathbb{N}$, the coefficients of the linear elliptic operator A are in $\mathcal{C}^{m+1}(\bar{U})$ and $\partial U \in \mathcal{C}^{m+2}$, then the unique solution $(u^N, \hat{u}^N)$ to (2.1)-(2.2) actually satisfies $(u^N, \hat{u}^N) \in H^{m+2}(U) \times H^{m+2}(U)$. In fact, if the coefficients of A are in $\mathcal{C}^\infty(\bar{U})$ and $\partial U \in \mathcal{C}^\infty$, then $(u^N, \hat{u}^N) \in \mathcal{C}^\infty(\bar{U}) \times \mathcal{C}^\infty(\bar{U})$. The interested reader is referred to classical manuscripts for more details, see for example [18, Chapter 8]. Here $\bar{U}$ is the closure of the set U and $\mathcal{C}^k$ denotes the set of $k-$times continuously differentiable functions.

Remark 2.5. The assumption of zero boundary conditions for the PDE (2.1)-(2.2) is for notational convenience and is done without loss of generality. For example, we can consider the PDE (2.1) with non-zero boundary data, say $u^N = q$, on ∂U, under the assumption $q \in H^1(U)$ for unique solvability of (2.1); see [18, Chapter 8] for more details.

Let us now describe the limiting system. For a given measure μ, let us set

$$B(x, x'; \mu) = \alpha \left\langle \sigma(wx' + \eta)\sigma(wx + \eta) + \sigma'(wx' + \eta)\sigma'(wx + \eta)c^2(x'x + 1), \mu \right\rangle. \tag{2.5}$$

Define the empirical measure

$$\mu_t^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{c_t^i, w_t^i, \eta_t^i}.$$

As it will be shown in the proof of Lemma 4.3, see equation (4.6), we can write the following representation for the neural network output

$$g_t^N(x) = g_0^N(x) - \int_0^t \int_U \hat{u}^N(s, x') B(x, x'; \mu_s^N) dx' ds. \tag{2.6}$$

Consider now the non-local PDE system

$$\begin{aligned} A u^*(t, x) &= g_t(x), \quad x \in U \\ u^*(t, x) &= 0, \quad x \in \partial U, \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} A^\dagger \hat{u}^*(t, x) &= (\langle u^*(t, x), m \rangle - \langle h, m \rangle)^\top m(x), \quad x \in U \\ \hat{u}^*(t, x) &= 0, \quad x \in \partial U. \end{aligned} \tag{2.8}$$

with

$$g_t(x) = -\int_0^t \left(\int_U \hat{u}^*(s, x') B(x, x'; \mu_0) dx' \right) ds. \tag{2.9}$$

The system of equations given by (2.7)-(2.8) is well-posed as our next Theorem 2.6 shows. Its proof is in Section 5.

Theorem 2.6. *Let as assume that Assumptions 2.1 and 2.2 hold. Then, there is a unique solution to the system (2.7)-(2.8) in* $C([0, \infty); H_0^1(U)) \times C([0, \infty); H_0^1(U))$.

We will show that $(u^N, \hat{u}^N, g^N) \to (u^*, \hat{u}^*, g)$ as $N \to \infty$. In particular, denoting by $\mathbb{E}$ the expectation operator associated with the randomness of the system (generated by the initial conditions), we have the following result whose proof is in Section 6.

Theorem 2.7. *Let $T > 0$ be given and assume the Assumptions 2.1, 2.2 and 2.3 hold. Then, as $N \to \infty$,*

$$\begin{aligned} \max_{0 \leqslant t \leqslant T} \mathbb{E} \left\| u^N(t) - u^*(t) \right\|_{H_0^1(U)} &\to 0, \\ \max_{0 \leqslant t \leqslant T} \mathbb{E} \left\| \hat{u}^N(t) - \hat{u}^*(t) \right\|_{H_0^1(U)} &\to 0, \\ \max_{0 \leqslant t \leqslant T} \mathbb{E} \left\| g_t^N - g_t \right\|_{L^2(U)} &\to 0. \end{aligned} \tag{2.10}$$

In addition, $\langle f, \mu_t^N \rangle \xrightarrow{p} \langle f, \mu_0 \rangle$ *(convergence in probability) for each $t \geqslant 0$ for all $f \in C_b^2(\mathbb{R}^{2+d})$.*

Let us return now to the objective function (1.8). $J_t^N \to \bar{J}_t$ as $N \to \infty$ where

$$\bar{J}_t = \frac{1}{2} \left\| \langle u^*(t) - h, m \rangle \right\|_2^2,$$

5

Differentiating this quantity yields

$$\frac{\partial \bar{J}_t}{\partial t} = \int_U \left(u^*(t,x) - h(x), m^\top \right) m(x) \frac{\partial u^*(t,x)}{\partial t} dx'. \tag{2.11}$$

Next notice that $\frac{\partial u^*(t,x)}{\partial t}$ is the solution to (2.7) with $f(x,\theta)$ replaced by $\dot{g}_t(x)$. Due to equation (2.9), we also know that

$$\dot{g}_t(x) = -\int_U \hat{u}^*(t,x') B(x,x';\mu_0) dx'.$$

Therefore,

$$\begin{aligned}
\frac{\partial \bar{J}_t}{\partial t} &= \int_U \left(u^*(t,x) - h(x), m^\top \right) m(x) \frac{\partial u^*(t,x)}{\partial t} dx' \\
&= \int_U A^\dagger \hat{u}^*(t,x) \frac{\partial u^*(t,x)}{\partial t} dx \\
&= \int_U \hat{u}^*(t,x) A \frac{\partial u^*(t,x)}{\partial t} dx \\
&= \int_U \hat{u}^*(t,x) \dot{g}_t(x) dx \\
&= -\int_U \hat{u}^*(t,x) \int_U \hat{u}^*(t,x') B(x,x';\mu_0) dx' dx \\
&= -\int_U \int_U \hat{u}^*(t,x) \hat{u}^*(t,x') B(x,x';\mu_0) dx' dx \\
&= -\int_\Theta \left| \int_U \hat{u}^*(t,x) \sigma(wx+\eta) dx \right|^2 d\mu_0(dcdwd\eta) - \int_\Theta \left| \int_U \hat{u}^*(t,x) c\sigma'(wx+\eta) x dx \right|^2 d\mu_0(dcdwd\eta) \\
&\quad - \int_\Theta \left| \int_U \hat{u}^*(t,x) c\sigma'(wx+\eta) dx \right|^2 d\mu_0(dcdwd\eta). \qquad\qquad \leqslant 0
\end{aligned} \tag{2.12}$$

The calculation (2.12) shows that the objective function $\bar{J}_t$ is monotonically decreasing. However, it is not immediately clear whether $u^*(t,\cdot)$ converges to a local minimum, saddle point, or global minimum. Using a spectral analysis, we are able to prove that $u^*(t,\cdot)$ will approach arbitrarily close to a global minimizer.

Theorem 2.8. *Let Assumptions 2.1, 2.2 and 2.3 hold. Then, for any $\epsilon > 0$ and $t > 0$, there exists a time $\tau > t$ such that*

$$\bar{J}_\tau < \epsilon.$$

Furthermore, we prove in Theorem 2.9 that the original pre-limit objective function J_t^N converges to a global minimum. Similar to the analysis in equation (2.12), one can also show that, for every $N \in \mathbb{N}$, J_t^N is non-increasing in t. Using (2.6) and a similar calculation as in (2.12),

$$\begin{aligned}
\frac{\partial J_t^N}{\partial t} &= -\int \left| \int_U \hat{u}^N(t,x) \sigma(wx+\eta) dx \right|^2 d\mu_t^N(dcdwd\eta) - \int \left| \int_U \hat{u}^N(t,x) c\sigma'(wx+\eta) x dx \right|^2 d\mu_t^N(dcdwd\eta) \\
&\quad - \int \left| \int_U \hat{u}^N(t,x) c\sigma'(wx+\eta) dx \right|^2 d\mu_t^N(dcdwd\eta) \\
&\leqslant 0
\end{aligned} \tag{2.13}$$

(2.13) will allow us to interchange the limits $N \to \infty$ and $t \to \infty$. This will be used to prove that $u^N(t,\cdot)$ converges to a global minimum.

Theorem 2.9. *Let Assumptions 2.1, 2.2 and 2.3 hold. For any $\epsilon > 0$, there exists an $N_0 \in \mathbb{N}$ such that*

$$\limsup_{t \to \infty} \sup_{N \geqslant N_0} \mathbb{E} J_t^N \leqslant \epsilon.$$

The proofs of Theorems 2.8 and 2.9 are in Sections 7 and 8.

Remark 2.10. In our theoretical analysis we have chosen the normalization of the neural network to be $1/N^\beta$ with $\beta \in (1/2, 1)$. We believe that similar results will also hold for the boundary cases $\beta = 1/2$ and $\beta = 1$, even though, the proofs presented below for the case $\beta \in (1/2, 1)$ suggest that if $\beta = 1/2$ or $\beta = 1$ additional terms survive in the limit that require extra analysis. In this first work on this topic we decided to present the proofs for $\beta \in (1/2, 1)$ and to leave the consideration of the boundary cases for future work.

3 Example Application: RANS Turbulence Closures

In this section, we demonstrate that the approach can also be applied in practice to nonlinear equations. We use a neural network for closure of the RANS equations, which leads to a nonlinear differential equation model with neural network terms. The neural network RANS model is then trained, using the adjoint equations, to match target data generated from the exact physics.

The incompressible Navier–Stokes equations

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho} \frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j} \left[\nu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right] \tag{3.1}$$

$$\frac{\partial u_j}{\partial x_j} = 0 \tag{3.2}$$

govern the evolution of a three-dimensional, unsteady velocity field $u_i(x_j, t) = (u, v, w)$, where $x_j = (x, y, z)$ are the Cartesian coordinates. Repeated indices imply summation. The mass density ρ and kinematic viscosity ν are constant, and the hydrodynamic pressure $p(x_j, t)$ is obtained from a Poisson equation [9] that discretely enforces the divergence-free constraint (3.2).

Direct Numerical Simulation (DNS) solves the Navier–Stokes equations with sufficient resolution such that the turbulence may be considered fully resolved. Doing so is prohibitively expensive for high Reynolds numbers or complicated geometries. The Reynolds-averaged Navier–Stokes (RANS) approach averages along statistically homogeneous directions (potentially including time) and so reduces the dimensionality of (3.1) and (3.2) with concomitant cost reduction. However, the RANS equations contain unclosed terms that require modeling.

3.1 Turbulent Channel Flow

An application to turbulent channel flow is considered here. The flow is statistically homogeneous in the streamwise direction (x), the spanwise direction (z), and time (t); hence the mean flow varies only in y. The RANS-averaged velocity components are

$$\overline{u}_i(y) = \frac{1}{L_x L_z (t_f - t_0)} \int_{t_0}^{t_f} \int_0^{L_z} \int_0^{L_x} u_i(x, y, z; t) \, \mathrm{d}x \, \mathrm{d}z \, \mathrm{d}t, \tag{3.3}$$

where L_x and L_z are the extents of the computational domain in the x- and z-directions, respectively, and $t_f - t_0$ is the time interval over which averaging is performed. The cross-stream and spanwise mean velocity components are identically zero: $\overline{v} = \overline{w} = 0$. A schematic of the flow is shown in Figure 1.

The three-dimensional, unsteady Navier–Stokes equations (3.1) and (3.2) are solved using DNS to provide high-resolution training and testing data. For turbulent channel flow, the characteristic length scale is the channel half-height $\delta = L_y/2$, and three characteristic velocity scales may be defined:

$$\overline{U} \equiv \frac{1}{\delta} \int_0^\delta \overline{u}(y) \, \mathrm{d}y \qquad \text{(bulk velocity)},$$

$$U_0 \equiv \overline{u}(y = 0) \qquad \text{(centerline velocity)},$$

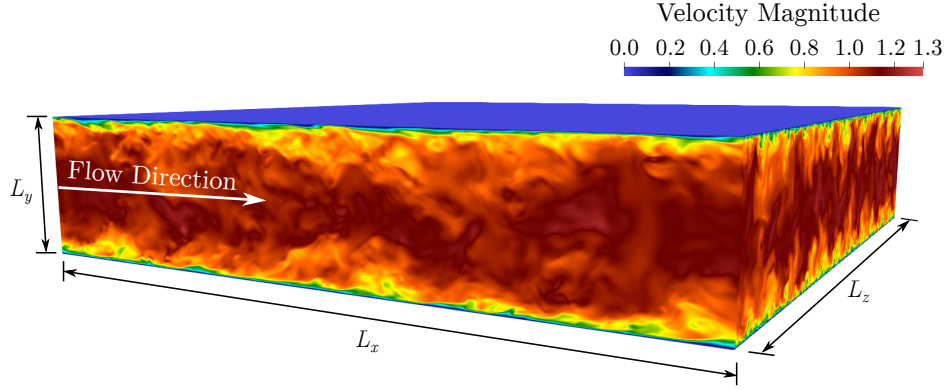$$u_\tau \equiv \sqrt{\tau_w / \rho} \qquad \text{(friction velocity)},$$

Figure 1: Schematic of the turbulent channel-flow configuration showing the domain extents and flow direction. Data is from instantaneous DNS of fully developed flow at Re $= 18 \times 10^3$ (Case D).

Case	Re	Re_0	Re_τ	C_f	Train	Test
A	6,001	3,380	190.0	8.02×10^{-3}	•	
B	9,000	5,232	270.3	7.22×10^{-3}		•
C	12,000	6,800	345.3	6.62×10^{-3}	•	
D	18,000	10,191	495.0	6.05×10^{-3}	•	
E	24,000	13,440	638.2	5.66×10^{-3}		•

Table 1: Reynolds numbers, friction coefficients, and train/test status for channel-flow DNS cases.

where $\tau(y) = \rho\nu(\mathrm{d}\overline{u}/\mathrm{d}y)$ is the mean shear stress and $\tau_w \equiv \tau(y = -\delta) = -\tau(y = \delta)$ is the wall shear stress. These define bulk, centerline, and friction Reynolds numbers

$$\mathrm{Re} \equiv \frac{2\delta\overline{U}}{\nu}, \qquad\qquad \mathrm{Re}_0 \equiv \frac{\delta U_0}{\nu}, \qquad\qquad \text{and} \qquad\qquad \mathrm{Re}_\tau \equiv \frac{u_\tau \delta}{\nu} = \frac{\delta}{\delta_v},$$

respectively, where $\delta_v = \nu/u_\tau$ is a viscous length scale. Flow in the near-wall boundary layer is typically presented in terms of "wall units" $y^+ = y/\delta_v$ and the corresponding normalized velocity $u^+ = \overline{u}/u_\tau$.

3.2 DNS Datasets

DNS datasets at Re $= \{6, 9, 12, 18, 24\} \times 10^3$ were generated for training and testing. Table 1 lists the bulk, centerline, and friction Reynolds numbers for each, as well as the computed friction coefficient

$$C_f \equiv \frac{\tau_w}{\frac{1}{2}\rho\overline{U}^2}.$$

The friction coefficient for the $\mathrm{Re}_\tau = 190$ case is in good agreement with that of the $\mathrm{Re}_\tau = 180$ DNS of Kim, Moin, and Moser [40]. Table 1 also lists the status of each dataset as it was used for training or testing. The near-wall streamwise mean velocity, in wall units, is shown for each case in Figure 2, showing good agreement with the theoretical log-law, particularly for the higher-Reynolds-number cases.

The DNS solves the Navier–Stokes equations using second-order centered differences on a staggered mesh [23]. Time integration is by a fractional-step method [39] with a linearized trapezoid method in an alternating-direction implicit (ADI) framework for the viscous terms [25]. A fixed time-step size is used, with resulting CFL numbers about 0.35.

Uniform mesh spacings $\Delta x = L_x/n_x$ and $\Delta z = L_z/n_z$ are used in the x- and z-directions, respectively, where n_x, n_y, and n_z are the number of mesh cells in each direction. The mesh is stretched in the y-direction,
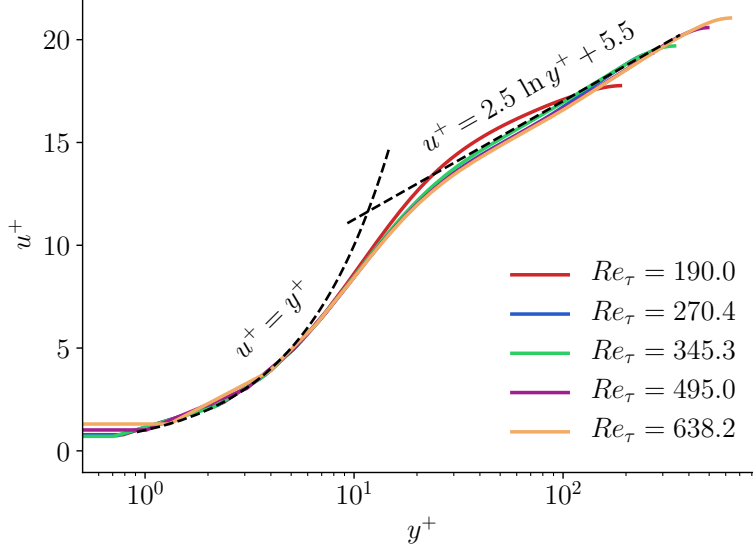
Figure 2: Mean velocity profiles in viscous units for DNS datasets listed in Table 1. The log-law is indicated by dashed lines for reference.

Case	$(L_x, L_y, L_z)/\delta$	$n_x \times n_y \times n_z$	Δy_w^+	Δy_0^+	Δx^+	Δz^+
A	$29 \times 2 \times 24$	$1024 \times 128 \times 768$	0.78	3.99	5.38	5.94
B	$10 \times 2 \times 10$	$1024 \times 192 \times 768$	0.74	3.78	2.64	3.52
C	$20 \times 2 \times 20$	$1280 \times 256 \times 1280$	0.71	3.61	5.39	5.39
D	$10 \times 2 \times 10$	$1280 \times 256 \times 1280$	1.02	5.19	3.87	3.87
E	$10 \times 2 \times 10$	$1280 \times 256 \times 1280$	1.30	6.67	4.99	4.99

Table 2: Domain sizes, grid sizes, and resolutions in wall units for turbulent channel-flow DNS cases.

with discrete cell-face points given by a hyperbolic tangent function

$$y_j = \delta \frac{\tanh\left[s_y\left(\frac{2j}{n_y} - 1\right)\right]}{\tanh s_y}, \qquad j = 0, 1, \ldots, n_y,$$

where $s_y = 1.05$ is a stretching parameter. Grid resolutions, listed in Table 2, are sufficient to place the first cell midpoint $0.71 \leqslant \Delta y_w^+ \leqslant 1.30$ from the wall with maximum centerline grid spacing $3.61 \leqslant \Delta y_0^+ \leqslant 6.67$. Depending on the case, the x- and z-directions are discretized with resolutions between $\Delta x^+ = \Delta x/\delta_v = 2.64$ and $\Delta z^+ = \Delta z/\delta_v = 5.39$.

3.3 Turbulent Channel-flow RANS

The momentum equation solved by RANS is obtained by applying (3.3) to (3.1) and letting $t_f \to \infty$. For turbulent channel flow, it reduces to a steady, one-dimensional equation for the streamwise mean velocity component $\overline{u}(y)$,

$$0 = -\frac{1}{\rho}\frac{\mathrm{d}\overline{p}}{\mathrm{d}x} + \frac{\mathrm{d}}{\mathrm{d}y}\left(\nu\frac{\mathrm{d}\overline{u}}{\mathrm{d}y} - \overline{u'v'}\right), \tag{3.4}$$

which is essentially the heat equation with a streamwise pressure-gradient forcing term $\mathrm{d}\overline{p}/\mathrm{d}x = -\tau_w/\delta$, which is adjusted to maintain constant $\overline{U} = 1$. Note that $\mathrm{d}\overline{p}/\mathrm{d}x$ is uniform in y. The term $\overline{u'v'}$ (the Reynolds stress) represents the action of the velocity fluctuations $u_i' = u_i - \overline{u}_i$ on the mean flow and is

unclosed. A standard closure model used for comparison is summarized in Sec. 3.3.1, and the neural network model to be tested is described in Sec. 3.3.2.

3.3.1   The k-ϵ Model

The eddy-viscosity (Boussinesq) hypothesis enables one to relate the Reynolds stress to the mean velocity gradients and so close (3.4). For turbulent channel flow, the eddy-viscosity hypothesis is stated

$$\overline{u'v'} = -\nu_t \frac{\mathrm{d}\overline{u}}{\mathrm{d}y}, \tag{3.5}$$

where ν_t is the eddy viscosity to be computed. Applying (3.5) to (3.4), the RANS equation is

$$0 = -\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x} + \frac{\mathrm{d}}{\mathrm{d}y}\left[(\nu + \nu_t)\frac{\mathrm{d}\overline{u}}{\mathrm{d}y}\right], \tag{3.6}$$

which is closed.

The eddy viscosity may be obtained in many ways. It is likely that the most common approach for engineering applications is the standard k-ϵ model [35, 45], which is a "two-equation" model in that it requires the solution of two additional PDEs. It computes ν_t from

$$\nu_t = C_\mu \frac{k^2}{\epsilon}, \tag{3.7}$$

where $k \equiv \frac{1}{2}\overline{u_i' u_i'}$ is the turbulent kinetic energy (TKE), ϵ is the TKE dissipation rate, and C_μ is a model parameter. The TKE and its dissipation rate may be computed from DNS and experimental data but are unclosed in RANS. Instead, the k-ϵ model obtains them from PDEs with modeled terms, which for channel flow reduce to

$$0 = \frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\nu + \nu_t}{\sigma_k}\frac{\mathrm{d}k}{\mathrm{d}y}\right) + \nu_t\left(\frac{\mathrm{d}\overline{u}}{\mathrm{d}y}\right)^2 - \epsilon,$$

$$0 = \frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\nu + \nu_t}{\sigma_\epsilon}\frac{\mathrm{d}\epsilon}{\mathrm{d}y}\right) + C_{1\epsilon}C_\mu k\left(\frac{\mathrm{d}\overline{u}}{\mathrm{d}y}\right)^2 - C_{2\epsilon}\frac{\epsilon^2}{k}. \tag{3.8}$$

Typical values for the constants are [45]:

$$C_\mu = 0.09, \qquad \sigma_k = 1.00, \qquad \sigma_\epsilon = 1.30, \qquad C_{1\epsilon} = 1.44, \quad \text{and} \quad C_{2\epsilon} = 1.92.$$

3.3.2   Neural Network Closure

Our neural network RANS model augments the eddy-viscosity model (3.5) by

$$\overline{u'v'} = -\nu_t\frac{\mathrm{d}\overline{u}}{\mathrm{d}y} + f_1\left(y, k, \epsilon, \frac{\mathrm{d}\overline{u}}{\mathrm{d}y}, \theta\right) + \frac{C_R}{\mathrm{Re}}f_2\left(y, k, \epsilon, \frac{\mathrm{d}\overline{u}}{\mathrm{d}y}, \theta\right), \tag{3.9}$$

where $f(\cdot, \theta)$ is a neural network with a vector output, $f_i(\cdot, \theta)$ is the i-th output of the neural network, and θ are its parameters. The constant $C_R = 6,001$ is the lowest Reynolds number in the training dataset. The input variables for the neural network are $(y, k, \epsilon, \mathrm{d}\overline{u}/\mathrm{d}y)$.

The RANS equation with the eddy-viscosity model and the neural network is

$$0 = -\frac{1}{\rho}\frac{\mathrm{d}\overline{p}}{\mathrm{d}x} + \frac{\mathrm{d}}{\mathrm{d}y}\left[(\nu + \nu_t)\frac{\mathrm{d}\overline{u}}{\mathrm{d}y} - f_1\left(y, k, \epsilon, \frac{\mathrm{d}\overline{u}}{\mathrm{d}y}, \theta\right) - \frac{C_R}{\mathrm{Re}}f_2\left(y, k, \epsilon, \frac{\mathrm{d}\overline{u}}{\mathrm{d}y}, \theta\right)\right], \tag{3.10}$$

with ν_t, k, and ϵ obtained as above.

The optimization of the PDE model (3.10) is a nonlinear version of our mathematical framework (1.2). The objective function for the model is the mean-squared error for the velocity profile (which can be thought of as a strong formulation of the objective function (1.1)):

$$J(\theta) = \frac{1}{2}\int_{-\delta}^{\delta}\left(h(y) - \overline{u}(y)\right)^2 dy, \tag{3.11}$$

10

where the target profile $h(y)$ is the averaged DNS data, and $\overline{u}$ is the solution to (3.10). The operator A from (1.2) has, in this case, become a nonlinear operator A acting on $(\overline{u}, k, \epsilon)$ and corresponding to the right-hand side of the equations (3.10) and (3.8). Specifically,

$$
A \begin{pmatrix} \overline{u} \\ k \\ \epsilon \end{pmatrix} = \begin{pmatrix} -\frac{1}{\rho}\frac{\mathrm{d}\overline{p}}{\mathrm{d}x} + \frac{\mathrm{d}}{\mathrm{d}y}[(\nu + \nu_t)\frac{\mathrm{d}\overline{u}}{\mathrm{d}y}] \\ \frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\nu+\nu_t}{\sigma_k}\frac{\mathrm{d}k}{\mathrm{d}y}\right) + \nu_t(\frac{\mathrm{d}\overline{u}}{\mathrm{d}y})^2 - \epsilon \\ \frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\nu+\nu_t}{\sigma_\epsilon}\frac{\mathrm{d}\epsilon}{\mathrm{d}y}\right) + C_{1\epsilon}C_\mu k(\frac{\mathrm{d}\overline{u}}{\mathrm{d}y})^2 - C_{2\epsilon}\frac{\epsilon^2}{k} \end{pmatrix},
$$

where we recall that the forcing term $\mathrm{d}\overline{p}/\mathrm{d}x = -\tau_w/\delta$ plus an adaptive term to enforce $\overline{U} \approx 1$, where $\tau_w = \rho\nu(\mathrm{d}\overline{u}/\mathrm{d}y)_{y=-\delta}$ is the wall shear stress defined in Subsection 3.1.

3.4   Numerical Results

The RANS equations are discretized using second-order centered differences on a staggered mesh. The solution $\overline{u}$ and its adjoint are computed at the y-faces, and the model variables k and ϵ are computed at cell centers. A coarse grid of $n_y = 32$ mesh points is used for all Reynolds numbers. The grid is nonuniform and is generated by the transformation

$$
\tilde{y} = \frac{\sin(\eta\pi\hat{y}/2)}{\sin(\eta\pi/2)}, \tag{3.12}
$$

where $\hat{y} \in [-\delta, \delta]$ is a uniform grid and $\eta = 0.995$. The system is solved using a pseudo-time-stepping scheme, in which the velocity solution is relaxed from an initially parabolic (laminar) profile using the fourth-order Runge–Kutta method until the solution has converged. Thus, although the RANS equations for channel flow satisfy a system of nonlinear differential equations, their solution is obtained by solving a time-dependent PDE.

   We train the neural network-augmented closure model (3.9) on the DNS-evaluated mean velocity for $\mathrm{Re} = \{6, 12, 18\} \times 10^3$. The adjoint of (3.10) is constructed using automatic differentiation as implemented by the Pytorch library [64]. 100 relaxation steps are taken per optimization iteration. Efficient, scalable training is performed using MPI parallelization and GPU offloading. The learning rate is decayed during training according to a piecewise constant learning rate schedule. Quadrature on the finite difference mesh is used to approximate the objective function.

   The neural network architecture is similar to that used for [49] and [62]. Exponential linear units (ELUs) and sigmoids are used for the activation functions. Note that ELUs violate the assumption on boundedness of the activation function under which our theoretical results hold. Nevertheless, this example demonstrates that the mathematical results of this paper may indeed be true under less stringent assumptions. The neural network architecture is:

$$
\begin{aligned}
H^{(1)}(z;\theta) &= \mathrm{ELU}(W^{(1)}z + b^{(1)}), \\
H^{(2)}(z;\theta) &= \mathrm{ELU}(W^{(2)}H^{(1)}(z;\theta) + b^{(2)}), \\
G(z;\theta) &= \sigma(W^{(3)}z + b^{(3)}), \\
U(z;\theta) &= G(z;\theta) \odot H^{(2)}(z;\theta), \\
f(z;\theta) &= W^{(4)}U(z;\theta) + b^{(4)},
\end{aligned} \tag{3.13}
$$

where the parameters $\theta = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(4)}\}$, $\odot$ denotes the element-wise product, $\mathrm{ELU}(\cdot)$ is an element-wise ELU activation function, and $\sigma(\cdot)$ is an element-wise sigmoid activation function.

   Once trained, the neural network-augmented model (3.9) is tested out-of-sample at $\mathrm{Re} = \{9, 24\} \times 10^3$. We observed that if the model was optimized for a very large number of iterations, instability could occur (presumably due to the equations becoming very stiff near the walls) or it could overfit on the test cases. The results presented are for a model trained with a moderately large number of optimization iterations which produces both good performances on the training cases as well as the test cases. The training mean-squared error is $\sim 1.8 \times 10^{-3}$. Figures 3 and 4 compare the mean velocity in mean-flow and wall units, respectively, for the DNS, RANS with the standard k-ϵ model, and RANS with the neural network-augmented
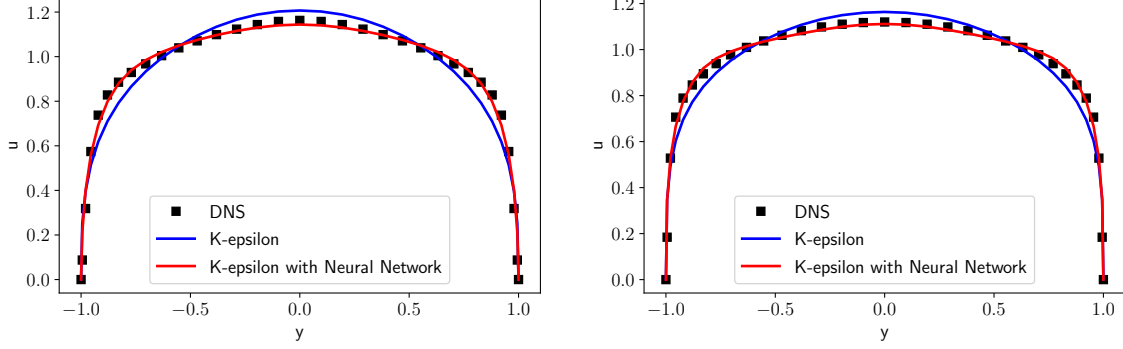
11

Figure 3: Mean velocity profiles obtained from spatio-temporally averaged DNS, RANS using the standard k-ϵ model, and RANS using the neural network-augmented model. DNS were computed using the meshes in Table 2; points indicate the cell-face locations used for the RANS calculations. Left: Re $= 9 \times 10^3$; right: Re $= 24 \times 10^3$. The neural network model was trained on Re $= \{6, 12, 18\} \times 10^3$.

model. The neural network model improves the accuracy of the k-ϵ model. In particular, the near-wall prediction (Figure 4) is improved by the neural network but still does not exactly match the DNS, as might be anticipated for the coarse grid.

4   A-priori preliminary bounds

Let us start by obtaining certain crucial a-priori bounds. The gradient of the objective function (1.1) can be evaluated using the solution $\hat{u}$ to the adjoint PDE (1.3). Indeed, this is the statement of Lemma 4.1 below.

Lemma 4.1. *Let Assumptions 2.1, 2.2 hold. Then,*

$$\nabla_\theta J(\theta) = \int_U \hat{u}(x; f) \nabla_\theta f(x, \theta) dx \tag{4.1}$$

Proof of Lemma 4.1. We start by defining $\tilde{u} = \nabla_\theta u$. If we differentiate (1.2) we obtain

$$A\tilde{u}(x) = \nabla_\theta f(x, \theta), \quad x \in U$$
$$\tilde{u}(x) = 0, \quad x \in \partial U$$

Integration by parts gives

$$(\hat{u}, A\tilde{u}) \quad = \quad (A^\dagger \hat{u}, \tilde{u}).$$

Using now (1.2) and (1.3) for the left and right hand side we obtain

$$(\langle u, m \rangle - \langle h, m \rangle)^\top (\tilde{u}, m) \quad = \quad (\hat{u}, \nabla_\theta f(\theta)).$$

Due to (1.1), we finally obtain

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \left[\frac{1}{2} (\langle u, m \rangle - \langle h, m \rangle)^\top (\langle u, m \rangle - \langle h, m \rangle) \right] \\
&= (\langle u, m \rangle - \langle h, m \rangle)^\top \langle \tilde{u}, m \rangle \\
&= (\hat{u}, \nabla_\theta f(\theta)),
\end{aligned} \tag{4.2}$$
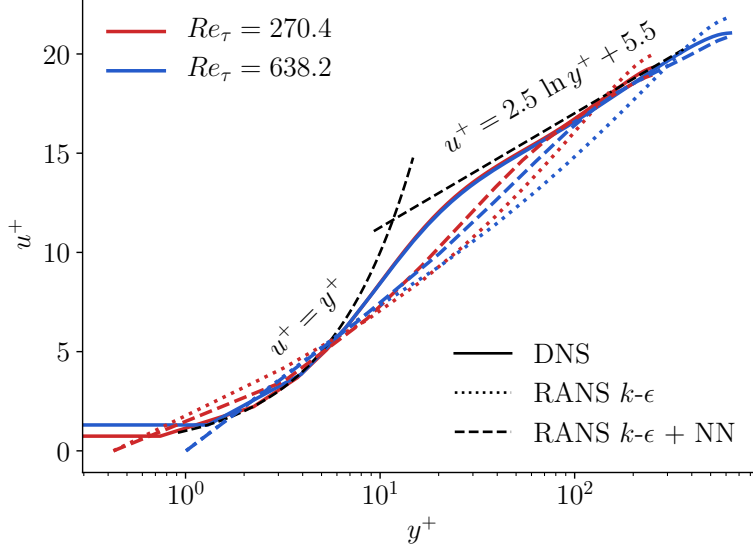
which concludes the proof of the lemma. $\qquad\square$

Figure 4: Mean velocity profiles in viscous units for out-of-sample testing cases: Re = $\{9, 24\} \times 10^3$. Solid lines: DNS; dotted lines: RANS with the k-ϵ model; dashed lines: RANS with the neural network-augmented model. The log-law is indicated by dashed lines for reference. The friction velocity used to normalize each Re_τ was computed from DNS. The solid lines for the two DNS datasets nearly overlap.

Lemma 4.1 shows that the gradient descent update can be written as follows:

$$
\begin{aligned}
\dot{\theta}_t^i &= -\alpha^N \nabla_{\theta_t^i} J(\theta_t) \\
&= -\alpha^{k,N} \left(\hat{u}^N(t,x), \nabla_{\theta_t^i} f(x, \theta_t) \right)_{L^2(U)},
\end{aligned}
\tag{4.3}
$$

It can be proven that the gradient descent updates are uniformly bounded with respect to t, N for $t \leqslant T$ for a given $T < \infty$.

Lemma 4.2. *Let $T < \infty$ be given and Assumptions 2.1-2.3 hold. There exists a finite constant $C < \infty$, depending on T and on the operator A, such that for all $i \in \mathbb{N}$ and all $t \in [0, T]$,*

$$
\sup_{i \in \{1, \cdots, N\}, N \in \mathbb{N}, t \in [0,T]} |c_t^i| \leqslant C < \infty,
$$

$$
\sup_{i \in \{1, \cdots, N\}, N \in \mathbb{N}, t \in [0,T]} \mathbb{E} \left(\left\| w_t^i \right\| + |\eta_t^i| \right) \leqslant C < \infty.
$$

Proof of Lemma 4.2. We start with a bound on $\left\| \hat{u}^N(t) \right\|_{H_0^1(U)}$. Using standard elliptic estimates based on the implied coercivity by Assumption 2.1 and the Cauchy-Schwarz inequality, we obtain that there is a constant $C < \infty$ such that

13

$$\left\|\hat{u}^N(t)\right\|_{H_0^1(U)}^2 \leqslant \left\|(\langle u^N(t), m\rangle - \langle h, m\rangle)^\top m\right\|_{L^2(U)}^2$$

$$\leqslant C \left\|\langle u^N(t), m\rangle - \langle h, m\rangle\right\|_2^2 \|m\|_{L^2(U)}^2$$

$$\leqslant C \left[\left\|u^N(t)\right\|_{L^2(U)}^2 + \|h\|_{L^2(U)}^2\right]$$

$$\leqslant C \left[\left\|u^N(t)\right\|_{H_0^1(U)}^2 + \|h\|_{L^2(U)}^2\right]$$

$$\leqslant C \left[\left\|g_t^N\right\|_{L^2(U)}^2 + \|h\|_{L^2(U)}^2\right]$$

$$\leqslant C \left[\frac{1}{N^{2\beta}} \int_U N \sum_{i=1}^N |c_t^i|^2 |\sigma(w_t^i x + \eta_t^i)|^2 dx + \|h\|_{L^2(U)}^2\right]$$

$$\leqslant C \left[\frac{1}{N^{2\beta-2}} \frac{1}{N} \sum_{i=1}^N |c_t^i|^2 + \|h\|_{L^2(U)}^2\right],$$

where in the last line we used the uniform boundedness of σ. Let us set $\kappa_t^N = \frac{1}{N} \sum_{i=1}^N |c_t^i|^2$. The last display gives us

$$\left\|\hat{u}^N(t)\right\|_{H_0^1(U)}^2 \leqslant C \left[\frac{1}{N^{2\beta-2}} \kappa_t^N + \|h\|_{L^2(U)}^2\right].$$

Next, we calculate

$$c_t^i = c_0^i - \frac{\alpha}{N^{1-\beta}} \int_0^t \int_U \hat{u}^N(s,x) \sigma(w_s^i x + \eta_s^i) dx ds,$$

which then implies due to the assumed uniform boundedness of σ that

$$|c_t^i| \leqslant |c_0^i| + \frac{\alpha \|\sigma\|}{N^{1-\beta}} \int_0^t \int_U |\hat{u}^N(s,x)| dx ds.$$

This then gives

$$|c_t^i|^2 \leqslant 2|c_0^i|^2 + \frac{C}{N^{2(1-\beta)}} t \int_0^t \left\|\hat{u}^N(s)\right\|_{L^2(U)}^2 ds$$

$$\leqslant 2|c_0^i|^2 + \frac{C}{N^{2(1-\beta)}} t \int_0^t \left[\frac{1}{N^{2\beta-2}} \kappa_s^N + \|h\|_{L^2(U)}^2\right] ds \tag{4.4}$$

Normalizing by N and summing up over all $i = 1, \cdots, N$ gives

$$\kappa_t^N \leqslant 2\kappa_0^N + \frac{C}{N^{2(1-\beta)}} t \int_0^t \left[\frac{1}{N^{2\beta-2}} \kappa_s^N + \|h\|_{L^2(U)}^2\right] ds$$

$$\leqslant 2\kappa_0^N + Ct \int_0^t \kappa_s^N ds + \frac{Ct^2}{N^{2-2\beta}} \|h\|_{L^2(U)}^2$$

Since c_0^i take values in a compact set, there exists a constant $C < \infty$ such that $\kappa_0^N \leqslant C$. Hence, using the discrete Gronwall lemma we get that there is a constant (still denoted by) C that may depend on T, but not on N, such that

$$\sup_{t \in [0,T]} \kappa_t^N \leqslant C.$$

Combining the latter with (4.4) gives for some unimportant constant $C < \infty$

$$|c_t^i|^2 \leqslant 2|c_0^i|^2 + C + \frac{C}{N^{2(1-\beta)}} \|h\|_{L^2(U)}^2, \tag{4.5}$$

14

from which the result follows, concluding the bound for $|c_t^i|$. Now, we turn to the bound for $\|w_t^i\|$. Using the uniform bound for $|C_k^i|$ and Assumptions 2.1-2.3 on the boundedness of the domain U and on σ' we get

$$\|w_t^i\| \leq \|w_0^i\| + \frac{\alpha}{N^{1-\beta}} \int_0^t \int_U |\hat{u}^N(s,x)||c_s^i|\sigma'(w_s^i x + \eta_s^i)x| dx ds$$

$$\leq \|w_0^i\| + \frac{C}{N^{1-\beta}} \int_0^t \|\hat{u}^N(s)\|_{L^1(U)} |c_s^i| ds$$

$$\leq \|w_0^i\| + \frac{C}{N^{1-\beta}} \int_0^t \left[\frac{1}{N^{\beta-1}} + \|h\|_{L^2(U)} \right] ds$$

$$\leq \|w_0^i\| + C \left[1 + \frac{1}{N^{1-\beta}} \right],$$

for a constant $C < \infty$ that may change from line to line. From this we directly obtain that

$$\sup_{i \in \{1,\cdots,N\}, N \in \mathbb{N}, t \in [0,T]} \mathbb{E} \|w_t^i\| \leq C < \infty.$$

The proof for the bound of $\sup_{i \in \{1,\cdots,N\}, N \in \mathbb{N}, t \in [0,T]} \mathbb{E}|\eta_t^i|$ is exactly the same as the uniform bound for $\mathbb{E} \|w_t^i\|$, concluding the proof of the lemma. $\qquad\square$

Next, in Lemma 4.3, we show an important a-priori bound relating the neural network output g^N with the solution to the adjoint equation $\hat{u}^N$.

Lemma 4.3. *For every $T < \infty$, we have that there is some constant $C < \infty$ (independent of N) such that for all $t \in [0,T]$ such that*

$$\max_{s \in [0,t]} \mathbb{E}\left[\|g_s^N\|_{L^2(U)}^2 \right] \leq C \left(1 + \int_0^t \max_{\tau \in [0,s]} \mathbb{E}\left[\|\hat{u}^N(\tau)\|_{L^2(U)}^2 \right] ds \right).$$

Proof of Lemma 4.3. Starting with $g_t^N(x) = \frac{1}{N^\beta} \sum_{i=1}^N c_t^i \sigma(w_t^i x + \eta_t^i)$ and differentiating using (2.3) we obtain

$$\dot{g}_t^N(x) = \frac{1}{N^\beta} \sum_{i=1}^N \left[\dot{c}_t^i \sigma(w_t^i x + \eta_t^i) + c_t^i \sigma'(w_t^i x + \eta_t^i)(\dot{w}_t^i x + \dot{\eta}_t^i) \right]$$

$$= -\frac{\alpha}{N} \sum_{i=1}^N \left[\int_U \hat{u}^N(t,x') \left(\sigma(w_t^i x + \eta_t^i)\sigma(w_t^i x' + \eta_t^i) + |c_t^i|^2 \sigma'(w_t^i x + \eta_t^i)\sigma'(w_t^i x' + \eta_t^i)(xx' + 1) \right) dx' \right]$$

Recalling now the empirical measure $\mu_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{c_t^i, w_t^i, \eta_t^i}$ and the definition of $B(x,x';\mu)$ by (2.5) we obtain the representation

$$\dot{g}_t^N(x) = -\int_U \hat{u}^N(t,x') B(x,x';\mu_s^N) dx' \tag{4.6}$$

The latter implies that

$$\|g_t^N\|_{L^2(U)}^2 \leq \|g_0^N\|_{L^2(U)}^2 + \int_U \left(\int_0^t \int_U \hat{u}^N(s,x') B(x,x';\mu_s^N) dx' ds \right)^2 dx$$

Notice now that for an appropriate constant C (independent of N) that may change from line to line

$$|B(x,x';\mu_s^N)| \leq \alpha \frac{1}{N} \sum_{i=1}^N \left| \sigma(w_s^i x + \eta_s^i)\sigma(w_s^i x' + \eta_s^i) + |c_s^i|^2 \sigma'(w_s^i x + \eta_s^i)\sigma'(w_s^i x' + \eta_s^i)(xx' + 1) \right|$$

$$\leq C \left[1 + \frac{1}{N} \sum_{i=1}^N |c_s^i|^2 \right]$$

$$\leq C,$$

15

due to the boudnedness of σ, σ', U and by Lemma 4.2. Thus, we have for an appropriate constant C that

$$\left\|g_t^N\right\|_{L^2(U)}^2 \leqslant \left\|g_0^N\right\|_{L^2(U)}^2 + C \int_U \left(\int_0^t \int_U |\hat{u}^N(s,x')| dx' ds\right)^2 dx$$

$$\leqslant \left\|g_0^N\right\|_{L^2(U)}^2 + C \left(\int_0^t \left\|\hat{u}^N(s)\right\|_{L^2(U)} ds\right)^2$$

Because now c_0^i, w_0^i, η_0^i are sampled independently, c_0^i take values in a compact set and σ is uniformly bounded we have

$$\mathbb{E}\left\|g_0^N\right\|_{L^2(U)}^2 = \frac{1}{N^{2\beta}} \int_U \mathbb{E}\left(\sum_{i=1}^N c_0^i \sigma(w_0^i x + \eta_0^i)\right)^2 dx$$

$$= \frac{1}{N^{2\beta}} \int_U \mathbb{E}\sum_{i=1}^N |c_0^i \sigma(w_0^i x + \eta_0^i)|^2 dx$$

$$\leqslant \frac{C}{N^{2\beta-1}}$$

$$\leqslant C, \tag{4.7}$$

where the last line follows because $\beta \in (1/2, 1)$.

Therefore, there is a constant C that is independent of N such that

$$\max_{s \in [0,t]} \mathbb{E}\left\|g_s^N\right\|_{L^2(U)}^2 \leqslant C\left(1 + \int_0^t \max_{\tau \in [0,s]} \mathbb{E}\left[\left\|\hat{u}^N(\tau)\right\|_{L^2(U)}^2\right] ds\right),$$

concluding the proof of the lemma.

$\square$

5 Well-posedness of limit PDE system, proof of Theorem 2.6

Let us define the linear operator $T_B : L^2(U) \to L^2(U)$ where

$$[T_B u](x) = \int_U B(x, x'; \mu_0) u(x') dx'.$$

The goal of this section is to prove well-posedness for the pair $(u^*, \hat{u}^*)$ solving the system of equations

$$Au^*(t, x) = g_t(x), \quad x \in U$$
$$u^*(t, x) = 0, \quad x \in \partial U, \tag{5.1}$$

and

$$A^\dagger \hat{u}^*(t, x) = (\langle u^*(t), m \rangle - \langle h, m \rangle)^\top m(x), \quad x \in U$$
$$\hat{u}^*(t, x) = 0, \quad x \in \partial U, \tag{5.2}$$

where

$$g_t(x) = -\int_0^t [T_B \hat{u}^*](s, x) ds \tag{5.3}$$

Proof of Theorem 2.6. We first prove that any solution to (5.1)-(5.2) is unique. Let's suppose there are two solutions $(u^{1,*}, \hat{u}^{1,*})$ and $(u^{2,*}, \hat{u}^{2,*})$. Let g^i be the solution to (5.3) corresponding to $\hat{u}^{i,*}$. Define

$$\phi(t, x) = u^{1,*}(t, x) - u^{2,*}(t, x),$$
$$\psi(t, x) = \hat{u}^{1,*}(t, x) - \hat{u}^{2,*}(t, x),$$
$$\zeta(t, x) = g_t^1(x) - g_t^2(x).$$

16

Then, (ϕ, ψ, ζ) satisfy

$$
\begin{aligned}
A\phi(t,x) &= \zeta(t,x), \\
A^\dagger \psi(t,x) &= (\langle \phi(t), m \rangle)^\top m(x), \\
\frac{d\zeta}{dt}(t,x) &= -[T_B \psi](t,x),
\end{aligned} \tag{5.4}
$$

with initial condition $\zeta(0,x) = 0$.

By standard elliptic systems made possible due to Assumption 2.1, we have the bounds

$$
\begin{aligned}
\|\phi(t)\|_{H_0^1(U)} &\leqslant C \|\zeta(t)\|_{L^2(U)}, \\
\|\psi(t)\|_{H_0^1(U)} &\leqslant C \|\phi(t)\|_{H_0^1(U)}.
\end{aligned} \tag{5.5}
$$

Let's now study the evolution of $\frac{1}{2}(\zeta(t), \zeta(t))$. Due to the bounbdedness of the domain U and consequently the boundedness of $B(x, x'; \mu_0)$ we obtain

$$
\begin{aligned}
\frac{d}{dt} \frac{1}{2} (\zeta(t), \zeta(t)) &= (\zeta(t), \frac{d\zeta}{dt}(t)) \\
&= -\int_{U \times U} \zeta(t,x) B(x, x'; \mu_0) \psi(t, x') dx' dx \\
&\leqslant C \|\zeta(t)\|_{L^2(U)} \|\psi(t)\|_{L^2(U)},
\end{aligned}
$$

where we have used the Cauchy-Schwarz inequality in the last line.

Using the inequality (5.5),

$$
\begin{aligned}
\frac{d}{dt} \frac{1}{2} (\zeta(t), \zeta(t)) &\leqslant C \|\zeta(t)\|_{L^2(U)} \|\psi(t)\|_{L^2(U)} \\
&\leqslant C \|\zeta(t)\|_{L^2(U)} \|\zeta(t)\|_{L^2(U)} \\
&= C(\zeta(t), \zeta(t)).
\end{aligned}
$$

By Gronwall's inequality, for $t \geqslant 0$,

$$
\|\zeta(t)\|_{L^2(U)} = 0.
$$

Due to the inequalities (5.5), for $t \geqslant 0$,

$$
\begin{aligned}
\|\phi(t)\|_{H_0^1(U)} &= 0, \\
\|\psi(t)\|_{H_0^1(U)} &= 0.
\end{aligned}
$$

Therefore, we have proven that any solution $(u^*, \hat{u}^*)$ to the PDE system (5.1)-(5.2) must be unique.

Let us now show existence. We will use a standard fixed point theorem argument. For given $T > 0$, let $U_T = U \times [0, T]$ and $V_T = C([0, T]; H_0^1(U))$ be the Banach space consisting of elements of $H_0^1(U_T)$ having a finite norm

$$
\|v\|_{V_T} = \max_{t \in [0,T]} \|v(t)\|_{H_0^1(U)}.
$$

Let $T > 0$ be given and $\ell \in L^2([0, T] \times U)$ be a given function. Consider the system of PDEs

$$
\begin{aligned}
Aw(t,x) &= -\int_0^t \ell(s,x) ds, \quad x \in U \\
w(t,x) &= 0, \quad x \in \partial U,
\end{aligned} \tag{5.6}
$$

and

$$
\begin{aligned}
A^\dagger \hat{w}(t,x) &= (\langle w(t) - h, m \rangle)^\top m(x), \quad x \in U \\
\hat{w}(t,x) &= 0, \quad x \in \partial U.
\end{aligned} \tag{5.7}
$$

17

Due to Assumption 2.1, the classical Lax-Milgram theorem guarantees that there is a solution system $(w(t), \hat{w}(t)) \in H_0^1(U) \times H_0^1(U)$ of (5.6)-(5.7) for all $t \in [0, T]$. As a matter of fact this means that $(w, \hat{w}) \in V_T \times V_T$.

Now for an arbitrarily given $\hat{u}^*(t) \in H_0^1(U)$, let us set $\ell(t, x) = [T_B \hat{u}^*](t, x)$. Notice that Cauchy-Schwarz inequality gives the estimate

$$\|\ell(t)\|_{L^2(U)} \leqslant \int_U ([T_B \hat{u}^*](t, x))^2 dx$$

$$\leqslant \int_U \left(\int_U B(x, x'; \mu_0) \hat{u}^*(t, x') dx' \right)^2 dx$$

$$\leqslant \|B(\cdot, \cdot; \mu_0)\|_{L^2(U \times U)} \|\hat{u}^*(t)\|_{L^2(U)}$$

$$\leqslant C \|\hat{u}^*(t)\|_{L^2(U)}$$

Consider now the map $L : V_T \times V_T \mapsto V_T \times V_T$ by setting $L(u^*, \hat{u}^*) = (w, \hat{w})$. For given $M < \infty$ and $T < \infty$, let us further define the space

$$V_T(M) = \{u \in V_T \text{ and } \|u\|_{V_T} \leqslant M\}.$$

We first show that L maps $V_T(M) \times V_T(M)$ into itself for appropriate choices of T and M. By standard elliptic estimates implied by Assumptions 2.1 and 2.2 we get for some unimportant constant $C_1 < \infty$ that

$$\|w\|_{V_T} + \|\hat{w}\|_{V_T} \leqslant C_1 T \left[\|u^*\|_{V_T} + \|\hat{u}^*\|_{V_T} \right] + C_1 \|h\|_{L^2(U)} \tag{5.8}$$

This means that if $u^*, \hat{u}^* \in V_T(M)$, then $w, \hat{w} \in V_T(M)$ if

$$2 C_1 T M + C_1 \|h\|_{L^2(U)} \leqslant M$$

Such a choice does exist and in particular we choose T, M so that $0 < T < \frac{M - C_1 \|h\|_{L^2(U)}}{2 C_1 M}$. Let's denote such a time T by $T_{1,1}$ and such $M = M_1$.

Next, we claim that if $T < \infty$ is small enough, then L is a contraction mapping. For this purpose, consider two pairs $(u_1^*, \hat{u}_1^*)$, $(u_2^*, \hat{u}_2^*)$ and their corresponding $(w_1, \hat{w}_1)$ and $(w_2, \hat{w}_2)$ through the mapping L. We have that

$$A(w_1 - w_2)(t, x) = - \int_0^t [T_B(\hat{u}_1^* - \hat{u}_2^*)](s, x) ds, \quad x \in U$$

$$w_1(t, x) = w_2(t, x) = 0, \quad x \in \partial U,$$

and

$$A^\dagger(\hat{w}_1 - \hat{w}_2)(t, x) = (\langle w_1(t) - w_2(t), m \rangle)^\top m(x), \quad x \in U$$

$$\hat{w}_1(t, x) = \hat{w}_2(t, x) = 0, \quad x \in \partial U.$$

By standard elliptic estimates implied by Assumptions 2.1 and 2.2 we get for some unimportant constant $C_2 < \infty$ that may change from line to line

$$\|\hat{w}_1 - \hat{w}_2\|_{V_T} \leqslant C_2 \|w_1 - w_2\|_V$$

$$\leqslant C_2 \int_0^T \|[T_B(\hat{u}_1^* - \hat{u}_2^*)](s)\|_{L_0^2(U)} ds$$

$$\leqslant C_2 \int_0^T \|(\hat{u}_1^* - \hat{u}_2^*)(s)\|_{L_0^2(U)} ds$$

$$\leqslant C_2 T \|(\hat{u}_1^* - \hat{u}_2^*)\|_{V_T}$$

The latter implies that there exists a constant $C_2 < \infty$ such that

$$\|w_1 - w_2\|_{V_T} + \|\hat{w}_1 - \hat{w}_2\|_{V_T} \leqslant C_2 T \left[\|u_1^* - u_2^*\|_{V_T} + \|\hat{u}_1^* - \hat{u}_2^*\|_{V_T} \right] \tag{5.9}$$

18

This means that the mapping L is indeed a contraction on the space $V_T \times V_T$ if T is chosen such that $C_2 T < 1$. Let's denote such a time T by $T_{1,2}$.

Let's set now $T_1 = \min\{T_{1,1}, T_{1,2}\}$. Hence, by Banach's fixed point theorem there is a solution $(u^*, \hat{u}^*)$ on $V_{T_1}(M_1) \times V_{T_1}(M_1)$.

Therefore, we have established the existence of a local in time solution to the system (5.1)-(5.2). Let us now extend this to a global solution in $[0, \infty)$. To do so we need to establish a global upper bound for the $\|\cdot\|_{H_0^1(U)} -$norm of solutions. Indeed, we first notice that

$$
\begin{aligned}
\|g_t\|_{L^2(U)}^2 &\leqslant \int_U \left(\int_0^t [T_B \hat{u}^*](s,x) ds \right)^2 dx \\
&\leqslant \int_U \left(\int_0^t \int_U B(x,x';\mu_0) \hat{u}^*(s,x') dx' ds \right)^2 dx \\
&\leqslant C \left(\int_0^t \int_U |\hat{u}^*(s,x')| dx' ds \right)^2 \\
&\leqslant C \left(\int_0^t \|\hat{u}^*(s)\|_{L^1(U)} ds \right)^2
\end{aligned}
$$

where we used the boundbedness of the kernel $B(x,x';\mu_0)$ and of U. From this we obtain that for some constant $C < \infty$

$$
\|g_t\|_{L^2(U)} \leqslant C \int_0^t \|\hat{u}^*(s)\|_{L^1(U)} ds
$$

Going back to the elliptic estimates guaranteed by Assumption 2.1 we subsequently obtain

$$
\begin{aligned}
\|\hat{u}^*(t)\|_{H_0^1(U)} &\leqslant C \left(\|u^*(t)\|_{H_0^1(U)} + \|h\|_{L^2(U)} \right) \\
&\leqslant C \left(\|g_t\|_{L^2(U)} + \|h\|_{L^2(U)} \right) \\
&\leqslant C \left(\int_0^t \|\hat{u}^*(s)\|_{L^1(U)} ds + \|h\|_{L^2(U)} \right) \\
&\leqslant C \left(\int_0^t \|\hat{u}^*(s)\|_{H_0^1(U)} ds + \|h\|_{L^2(U)} \right)
\end{aligned}
$$

By Grownall lemma now we have

$$
\|\hat{u}^*(t)\|_{H_0^1(U)} \leqslant \|h\|_{L^2(U)} e^{Ct},
$$

and as a matter of fact we have the global upper bound for some constants $C_1, C_2 < \infty$

$$
\|u^*(t)\|_{H_0^1(U)} + \|\hat{u}^*(t)\|_{H_0^1(U)} \leqslant C_1 \|h\|_{L^2(U)} \left(e^{C_2 t} + 1 \right).
$$

Notice that this a-priori global upper bound immediately implies that if the maximal interval of existence of the solution $(u^*, \hat{u}^*)$, say $[0, T^*)$ were to be bounded, i.e., if $T^* < \infty$, then $(u^*, \hat{u}^*)$ would be uniformly bounded in $H_0^1(U)$ in $[0, T^*]$ and would not explode in finite time, see [1, 46].

With this a-priori bound at hand the solution can be extended indefinitely in time. Indeed, one can proceed now in a standard way, see for example the standard iterative argument in Section 9.2 of [16], as well as the proof of Proposition 6.1 of [46] for a setting closely related to ours and Section 5 of [1] for related results. Thus a unique global solution in $C([0, \infty); H_0^1(U)) \times C([0, \infty); H_0^1(U))$ exists, concluding the proof of the theorem.

$\square$

6   Proof of Theorem 2.7

We will now show that the solution to (2.1)-(2.2) converges to the solution to (2.7)-(2.8) as $N \to \infty$. We start with the following uniform boundedness lemma with respect to N for the system (2.1)-(2.2).

Lemma 6.1. *Let $T > 0$ be given and assume the Assumption 2.1 holds. Then there exists a constant $C < \infty$ independent of N such that*

$$\max_{t \in [0,T]} \mathbb{E}\left[\left\| u^N(t) \right\|_{H_0^1(U)}^2 \right] + \max_{t \in [0,T]} \mathbb{E}\left[\left\| \hat{u}^N(t) \right\|_{H_0^1(U)}^2 \right] \leqslant C.$$

Proof of Lemma 6.1. By standard energy estimates based on the coercivity property implied by Assumption 2.1 we have that there is a constant C_0 that may change from line to line such that

$$
\begin{aligned}
\max_{s \in [0,t]} \mathbb{E}\left\| \hat{u}^N(t) \right\|_{H_0^1(U)}^2 & \leqslant C_0 \left[\max_{s \in [0,t]} \mathbb{E}\left\| u^N(s) \right\|_{H_0^1(U)}^2 + \|h\|_{L^2(U)}^2 \right] \\
& \leqslant C_0 \left[\max_{s \in [0,t]} \mathbb{E}\left\| g_s^N \right\|_{L^2(U)}^2 + \|h\|_{L^2(U)}^2 \right] \\
& \leqslant C_0 \left[1 + \int_0^t \max_{\tau \in [0,s]} \mathbb{E}\left\| \hat{u}^N(\tau) \right\|_{H_0^1(U)}^2 ds + \|h\|_{L^2(U)}^2 \right].
\end{aligned}
\tag{6.1}
$$

where the third line uses Lemma 4.3. The bound for $\max_{t \in [0,T]} \mathbb{E}\left\| \hat{u}^N(t) \right\|_{H_0^1(U)}^2$ then follows by Gronwall's lemma using the fact that, by assumption, $h \in L^2$. The bound for $\max_{t \in [0,T]} \mathbb{E}\left\| u^N(t) \right\|_{H_0^1(U)}^2$ follows directly from equation (6.1) and the bound for $\max_{t \in [0,T]} \mathbb{E}\left\| \hat{u}^N(t) \right\|_{H_0^1(U)}^2$. This concludes the proof of the lemma. $\square$

Now we can proceed with the proof of Theorem 2.7.

Proof of Theorem 2.7. Recall that the parameters c_t^i satisfy the equation:

$$c_t^i = c_0^i - \frac{\alpha}{N^{1-\beta}} \int_0^t \int_U \hat{u}^N(s,x) \sigma(w_s^i x + \eta_s^i) dx ds.$$

Re-arranging and using the Cauchy-Schwarz inequality yields:

$$
\begin{aligned}
|c_t^i - c_0^i| & \leqslant \frac{\alpha}{N^{1-\beta}} \int_0^t \int_U |\hat{u}^N(s,x)| |\sigma(w_s^i x + \eta_s^i)| dx ds \\
& \leqslant \frac{C}{N^{1-\beta}} \int_0^t \int_U |\hat{u}^N(s,x)| dx ds \\
& \leqslant \frac{C}{N^{1-\beta}} \int_0^t \left(\int_U |\hat{u}^N(s,x)|^2 dx \right)^{\frac{1}{2}} ds \\
& = \frac{C}{N^{1-\beta}} \int_0^t \left\| \hat{u}^N(s) \right\|_{L^2(U)} ds.
\end{aligned}
\tag{6.2}
$$

Taking expectations of both sides of the above equation leads to the following uniform bound:

$$
\begin{aligned}
\sup_{i \in \{1,\dots,N\}, t \in [0,T]} \mathbb{E}|c_t^i - c_0^i| & \leqslant \frac{C}{N^{1-\beta}} \int_0^T \mathbb{E}\left\| \hat{u}^N(s) \right\|_{L^2(U)} ds \\
& \leqslant \frac{C}{N^{1-\beta}}.
\end{aligned}
\tag{6.3}
$$

Let us now consider the parameters w_t^i. Using the uniform bound $\sup\limits_{N\in\mathbb{N},i\in\{1,\dots,N\},t\in[0,T]}|c_t^i| \leqslant C$ from Lemma 4.2 and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
|w_t^i - w_0^i| &\leqslant \frac{C}{N^{1-\beta}} \int_0^t \int_U |\hat{u}^N(s,x)||c_s^i|\sigma'(w_s^i x + \eta_s^i)x|dxds, \\
&\leqslant \frac{C}{N^{1-\beta}} \int_0^t \int_U |\hat{u}^N(s,x)|dxds \\
&\leqslant \frac{C}{N^{1-\beta}} \int_0^t \left\|\hat{u}^N(s)\right\|_{L^2(U)} ds.
\end{aligned}
$$
(6.4)

Taking expectations, we have the uniform bound

$$
\sup_{i\in\{1,\dots,N\},t\in[0,T]} \mathbb{E}|w_t^i - w_0^i| \leqslant \frac{C}{N^{1-\beta}}.
$$
(6.5)

A similar uniform bound holds for the parameters η_t^i.

Let $f \in C_b^2(\mathbb{R}^{2+d})$. Then, using the global Lipschitz property of f,

$$
\begin{aligned}
|\langle f, \mu_t^N - \mu_0\rangle| &\leqslant |\langle f, \mu_t^N - \mu_0^N\rangle| + |\langle f, \mu_0^N - \mu_0\rangle| \\
&\leqslant \frac{C}{N} \sum_{i=1}^N \left(|w_t^i - w_0^i| + |c_t^i - c_0^i| + |\eta_t^i - \eta_0^i|\right) + |\langle f, \mu_0^N - \mu_0\rangle|.
\end{aligned}
$$

Combining the uniform bounds with the equation above, taking expectations, using Markov's inequality and the fact that $\mu_0^N \to \mu_0$ as $N \to \infty$, proves that $\langle f, \mu_t^N\rangle \xrightarrow{p} \langle f, \mu_0\rangle$ as $N \to \infty$ for each $t \geqslant 0$.

Let us next turn our attention to (9.7). Due to the boundedness of U and of σ, σ', we obtain

$$
\left\|B(\cdot,\cdot;\mu_s^N)\right\|_{L^2(U\times U)} \leqslant C\left[1 + \frac{1}{N}\sum_{i=1}^N |c_s^i|^2\right]
$$

By Lemma 4.2 we then have that there is a universal constant $C < \infty$ such that

$$
\sup_{s\in[0,T]} \left\|B(\cdot,\cdot;\mu_s^N)\right\|_{L^2(U\times U)} \leqslant C
$$
(6.6)

Let $(u^*, \hat{u}^*)$ be the unique solution to the system given in Theorem 2.6. Let us define

$$
\begin{aligned}
\phi^N(t,x) &= u^N(t,x) - u^*(t,x), \\
\psi^N(t,x) &= \hat{u}^N(t,x) - \hat{u}^*(t,x), \\
\zeta^N(t,x) &= g_t^N(x) - g_t(x).
\end{aligned}
$$

We recall from (4.6) the representation

$$
g_t^N(x) = g_0^N(x) - \int_0^t \left(\int_U \hat{u}^N(s,x')B(x,x';\mu_s^N)dx'\right) ds
$$
(6.7)

Then, we write

$$
\begin{aligned}
\zeta^N(t,x) &= g_t^N(x) - g_t(x) \\
&= \zeta^N(0,x) - \int_0^t \int_U \hat{u}^N(s,x')\left(B(x,x';\mu_s^N) - B(x,x';\mu_0)\right) dx'ds \\
&\quad - \int_0^t \int_U \psi^N(s,x')B(x,x';\mu_0)dx'ds
\end{aligned}
$$

21

By taking L^2 bounds we obtain

$$\left\|\zeta^N(t)\right\|_{L^2(U)} \leqslant \left\|\zeta^N(0)\right\|_{L^2(U)} + \int_0^t \left\|\hat{u}^N(s)\right\|_{L^2(U)} \left\|B(\cdot,\cdot;\mu_s^N - \mu_0)\right\|_{L^2(U \times U)} ds$$

$$+ \int_0^t \left\|\psi^N(s)\right\|_{L^2(U)} \left\|B(\cdot,\cdot;\mu_0)\right\|_{L^2(U \times U)} ds \qquad (6.8)$$

Standard elliptic PDE estimates based on coercivity, the Cauchy-Schwarz inequality, and taking expectations yields the following bound:

$$
\begin{aligned}
\mathbb{E}\left\|\psi^N(t)\right\|_{H_0^1(U)} &\leqslant \mathbb{E}\left\|\psi^N(t)\right\|_{L^2(U)} \\
&= \mathbb{E}\left\|\left\langle\phi^N(t), m\right\rangle^\top m\right\|_{L^2(U)} \\
&\leqslant C\mathbb{E}\left\|\phi^N(t)\right\|_{L^2(U)} \\
&\leqslant C\mathbb{E}\left\|\phi^N(t)\right\|_{H_0^1(U)} \\
&\leqslant C_0\mathbb{E}\left\|\zeta^N(t)\right\|_{L^2(U)} \\
&\leqslant C\left(\mathbb{E}\left\|\zeta^N(0)\right\|_{L^2(U)} + \int_0^t \mathbb{E}\left[\left\|\hat{u}^N(s)\right\|_{L^2(U)} \left\|B(\cdot,\cdot;\mu_s^N - \mu_0)\right\|_{L^2(U \times U)}\right] ds \right. \\
&\quad + \left. \int_0^t \mathbb{E}\left\|\psi^N(s)\right\|_{H_0^1(U)} \left\|B(\cdot,\cdot;\mu_0)\right\|_{L^2(U \times U)} ds\right). \qquad (6.9)
\end{aligned}
$$

Let us now examine the second term on the second line of the above equation. By the Cauchy-Schwarz inequality and Lemma 6.1,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{u}^N(s)\right\|_{L^2(U)} \left\|B(\cdot,\cdot;\mu_s^N - \mu_0)\right\|_{L^2(U \times U)}\right] &\leqslant \mathbb{E}\left[\left\|\hat{u}^N(s)\right\|_{L^2(U)}^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left\|B(\cdot,\cdot;\mu_s^N - \mu_0)\right\|_{L^2(U \times U)}^2\right]^{\frac{1}{2}} \\
&\leqslant C\mathbb{E}\left[\left\|B(\cdot,\cdot;\mu_s^N - \mu_0)\right\|_{L^2(U \times U)}^2\right]^{\frac{1}{2}}.
\end{aligned}
$$

Let $f(x,x',c,w,\eta) = \alpha\sigma(wx'+\eta)\sigma(wx+\eta) + \sigma'(wx'+\eta)\sigma'(wx+\eta)c^2(x'x+1)$. Using a Taylor expansion, the fact that U is a bounded set, and the uniform bound on c_t^i produces

$$
\begin{aligned}
\left\|B(\cdot,\cdot;\mu_s^N - \mu_0)\right\|_{L^2(U \times U)}^2 &= \int_U \int_U \left\langle f(x,x',\cdot), \mu_s^N - \mu_0\right\rangle^2 dx dx' \\
&= \int_U \int_U \left\langle f(x,x',\cdot), \mu_s^N - \mu_0^N\right\rangle^2 dx dx' + \int_U \int_U \left\langle f(x,x',\cdot), \mu_0^N - \mu_0\right\rangle^2 dx dx'. \\
&= \int_U \int_U \frac{C}{N^2}\left(\sum_{i=1}^N [|w_s^i - w_0^i| + |c_s^i - c_0^i| + |\eta_s^i - \eta_0^i|]\right)^2 dx dx' \\
&\quad + \int_U \int_U \left(\frac{1}{N}\sum_{i=1}^N (f(x,x',c_0^i,w_0^i,\eta_0^i) - \left\langle f(x,x',\cdot), \mu_0\right\rangle\right)^2.
\end{aligned}
$$

The expectation of the last term is bounded since (c_0^i, w_0^i, η_0^i) are i.i.d. samples from μ_0. The Cauchy-Schwarz inequality and the inequalities (6.2) and (6.4) yield:

$$
\begin{aligned}
\int_U \int_U \frac{C}{N^2}\left(\sum_{i=1}^N [|w_s^i - w_0^i| + |c_s^i - c_0^i| + |\eta_s^i - \eta_0^i|]\right)^2 dx dx' &\leqslant \int_U \int_U \frac{C}{N^2}\left(\sum_{i=1}^N \frac{C}{N^{1-\beta}}\int_0^T \left\|\hat{u}^N(t)\right\|_{L^2(U)} dt\right)^2 dx dx' \\
&\leqslant \frac{C}{N^{2-2\beta}} \times \left(\int_0^T \left\|\hat{u}^N(s)\right\|_{L^2(U)} ds\right)^2 \\
&\leqslant \frac{C}{N^{2-2\beta}}\int_0^T \left\|\hat{u}^N(s)\right\|_{L^2(U)}^2 ds.
\end{aligned}
$$

22

Therefore, we get

$$\mathbb{E}\left[\left\| B(\cdot,\cdot; \mu_s^N - \mu_0) \right\|_{L^2(U \times U)}^2 \right] \;\leqslant\; \frac{C}{N^{2-2\beta}} \int_0^T \mathbb{E}\left[\left\| \hat{u}^N(s) \right\|_{L^2(U)}^2 \right] ds + \frac{C}{N}.$$

Combining these estimates, we now have the bound

$$\mathbb{E}\left\| \psi^N(t) \right\|_{H_0^1(U)} \;\leqslant\; C\left(\mathbb{E}\left\| \zeta^N(0) \right\|_{L^2(U)} + \frac{C}{N} + \int_0^t \mathbb{E}\left\| \psi^N(s) \right\|_{H_0^1(U)} ds \right).$$

This bound along with $\left\| \zeta^N(0) \right\|_{L^2(U)} \xrightarrow{p} 0$ (recall by (4.7) that $\mathbb{E}\left\| g_0^N \right\|_{L^2(U)}^2 \leqslant \frac{C}{N^{2\beta-1}} \to 0$) allow us to conclude via Gronwall's lemma that

$$\max_{t \in [0,T]} \mathbb{E}\left\| \psi^N(t) \right\|_{H_0^1(U)} \to 0, \text{ as } N \to \infty.$$

The latter then implies that

$$\max_{t \in [0,T]} \mathbb{E}\left\| \zeta^N(t) \right\|_{L^2(U)} \to 0, \text{ as } N \to \infty.$$

Consequently, elliptic estimates based on the coercivity of the operator A give

$$\max_{t \in [0,T]} \mathbb{E}\left\| \phi^N(t) \right\|_{H_0^1(U)} \leqslant C \max_{t \in [0,T]} \mathbb{E}\left\| \zeta^N(t) \right\|_{L^2(U)} \to 0, \text{ as } N \to \infty,$$

concluding the proof of the theorem. $\qquad\square$

7 Properties of the limiting system and related preliminary calculations

Let us define the linear operator $T_B : L^2(U) \to L^2(U)$ where

$$[T_B u](x) = \int_U B(x, x'; \mu_0) u(x') dx'.$$

Due to the fact that the kernel B is symmetric and of Hilbert-Schmidt type, we immediately get that the integral operator T_B is a self-adjoint, compact linear operator whose eigenfunctions $\{e_k\}_{k=1}^{\infty}$ form an orthonormal basis for $L^2(U)$, see for example Chapter 8.5 in [58].

Recall our objective function $\bar{J}(t) = \frac{1}{2} \langle u^*(t) - h, m \rangle^\top \langle u^*(t) - h, m \rangle$. Differentiating with respect to time yields

$$\begin{aligned}
\frac{d\bar{J}}{dt} &= \langle u^*(t) - h, m \rangle^\top \left\langle \frac{du^*}{dt}, m \right\rangle \\
&= \int_U \frac{du^*}{dt} \langle u^*(t) - h, m \rangle^\top m(x) dx \\
&= \int_U \frac{du^*}{dt} A^\dagger \hat{u}^* dx \\
&= \int_U \frac{dAu^*}{dt} \hat{u}^* dx \\
&= \int_U \frac{dg}{dt} \hat{u}^* dx \\
&= -\int_U \hat{u}^*(x) \int_U B(x, x') \hat{u}^*(x') dx' dx \\
&= -(\hat{u}^*, T_B \hat{u}^*).
\end{aligned} \tag{7.1}$$

Notice that

$$\frac{1}{2}(A^\dagger \hat{u}^*(t), A^\dagger \hat{u}^*(t)) = \frac{1}{2}(\langle u^*(t) - h, m\rangle^\top m, \langle u^*(t) - h, m\rangle^\top m)$$

$$= \frac{1}{2}\langle u^*(t) - h, m\rangle^\top (m, m^\top)\langle u^*(t) - h, m\rangle. \tag{7.2}$$

Differentiating with respect to time yields

$$\frac{d}{dt}(A^\dagger \hat{u}^*(t), A^\dagger \hat{u}^*(t)) = \frac{d}{dt}\langle u^*(t) - h, m\rangle^\top (m, m^\top)\langle u^*(t) - h, m\rangle. \tag{7.3}$$

By Assumption 2.2, we get due to the orthonormal assumption on the elements m_ℓ of the vector $\{m_\ell\}_{\ell=1}^L$

$$\frac{d}{dt}(A^\dagger \hat{u}^*(t), A^\dagger \hat{u}^*(t)) = \frac{d}{dt}\langle u^*(t) - h, m\rangle^\top I \langle u^*(t) - h, m\rangle$$

$$= \frac{d\bar{J}}{dt}$$

$$= -(\hat{u}^*, T_B \hat{u}^*)$$

$$\leqslant 0 \tag{7.4}$$

Therefore, for all $t \geqslant 0$ we will have

$$\frac{1}{2}(A^\dagger \hat{u}^*(t), A^\dagger \hat{u}^*(t)) = \frac{1}{2}(A^\dagger \hat{u}^*(0), A^\dagger \hat{u}^*(0)) - \int_0^t (\hat{u}^*(s), [T_B \hat{u}^*](s))ds. \tag{7.5}$$

Then, assuming $AA^\dagger$ is coercive,

$$(A^\dagger \hat{u}^*(t), A^\dagger \hat{u}^*(t)) = (\hat{u}^*(t), AA^\dagger \hat{u}^*(t))$$

$$\geqslant \gamma(\hat{u}^*(t), \hat{u}^*(t)),$$

where $\gamma > 0$.

Note that this establishes a uniform bound for $\|\hat{u}^*(t)\|^2$.

$$(\hat{u}^*(t), \hat{u}^*(t)) \leqslant \frac{1}{\gamma}(A^\dagger \hat{u}^*(t), A^\dagger \hat{u}^*(t))$$

$$\leqslant \frac{1}{\gamma}(A^\dagger \hat{u}^*(0), A^\dagger \hat{u}^*(0))$$

$$\leqslant C < \infty. \tag{7.6}$$

Note that the constant C does not depend upon time, and therefore

$$\sup_{t \geqslant 0} \|\hat{u}^*(t)\| < C. \tag{7.7}$$

The coercivity condition for $AA^\dagger$ holds if A is self-adjoint. Indeed, let us use the representation

$$\hat{u}^*(t, x) = \sum_{k=1}^\infty c_k(t)v_k(x),$$

where $v_k(x)$ are the eigenfunctions of A and form an orthonormal basis. We know that the corresponding eigenvalues $\lambda_k \geqslant \lambda > 0$. We will prove the coercivity condition when A is self-adjoint, i.e. $A^\dagger = A$. Define

$$\hat{u}_K^*(t, x) = \sum_{k=1}^K c_k(t)v_k(x).$$

Then,

$$\lim_{K \to \infty}(\hat{u}^*, AA^\dagger \hat{u}_K^*) = \lim_{K \to \infty}(AA^\dagger \hat{u}^*, \hat{u}_K^*). \tag{7.8}$$

24

Using the Cauchy-Schwarz inequality,

$$
\begin{aligned}
|(AA^\dagger \hat{u}^*, \hat{u}_K^*) - (AA^\dagger \hat{u}^*, \hat{u}^*)| & \leqslant |(|AA^\dagger \hat{u}^*|, |\hat{u}_K^* - \hat{u}^*|)| \\
& \leqslant \|AA^\dagger \hat{u}^*\| \, \|\hat{u}_K^* - \hat{u}^*\| .
\end{aligned}
$$

Note that

$$
AA^\dagger \hat{u}^* = A(\langle u - h, m \rangle)^\top m
$$

Consequently, $AA^\dagger \hat{u}^*$ is in L^2 as long as Am_ℓ are in L^2 (which is assumed in Assumption 2.2).
Since $\lim_{K \to \infty} \|\hat{u}_K^* - \hat{u}^*\|_{L^2(U)} = 0$,

$$
\lim_{K \to \infty} |(AA^\dagger \hat{u}^*, \hat{u}_K^*) - (AA^\dagger \hat{u}^*, \hat{u}^*)| = 0.
$$

Therefore,

$$
\begin{aligned}
\lim_{K \to \infty} (\hat{u}^*, AA^\dagger \hat{u}_K^*) & = \lim_{K \to \infty} (AA^\dagger \hat{u}^*, \hat{u}_K^*) \\
& = (AA^\dagger \hat{u}^*, \hat{u}^*).
\end{aligned}
\tag{7.9}
$$

On the other hand,

$$
\begin{aligned}
(\hat{u}^*(t), AA^\dagger \hat{u}_K^*(t)) & = (\hat{u}^*(t), AA^\dagger \sum_{k=1}^{K} c_k(t) v_k(x)) \\
& = (\hat{u}^*(t), \sum_{k=1}^{K} c_k(t) \lambda_k^2 v_k(x)) \\
& = (\sum_{k=1}^{\infty} c_k(t) v_k(x), \sum_{k=1}^{K} c_k(t) \lambda_k^2 v_k(x)) \\
& = \sum_{k=1}^{K} c_k^2(t) \lambda_k^2.
\end{aligned}
$$

From equation (7.9), we know that

$$
\lim_{K \to \infty} (\hat{u}^*(t), AA^\dagger \hat{u}_K^*(t)) = (\hat{u}^*(t), AA^\dagger \hat{u}^*(t)).
$$

Since it is also true that $(\hat{u}^*(t), AA^\dagger \hat{u}_K^*(t)) = \sum_{k=1}^{K} c_k^2(t) \lambda_k^2$,

$$
\lim_{K \to \infty} \sum_{k=1}^{K} c_k^2(t) \lambda_k^2 = (\hat{u}^*(t), AA^\dagger \hat{u}^*(t)).
$$

Consequently, since $\lambda_k \geqslant \lambda$,

$$
(\hat{u}^*(t), AA^\dagger \hat{u}^*(t)) \geqslant \lambda^2 \sum_{k=1}^{\infty} c_k^2(t).
$$

Finally, we know that

$$
\begin{aligned}
(\hat{u}^*(t), \hat{u}^*(t)) & = (\sum_{k=1}^{\infty} c_k(t) v_k(x), \sum_{k=1}^{\infty} c_k(t) v_k(x)) \\
& = \sum_{k=1}^{\infty} c_k^2(t).
\end{aligned}
$$

25

Therefore,

$$(\hat{u}^*(t), AA^\dagger \hat{u}^*(t)) \geqslant \lambda^2 (\hat{u}^*(t), \hat{u}^*(t)),$$

which proves the coercivity condition with $\gamma = \lambda^2$ when A is self-adjoint.

Therefore, using the coercivity property of $AA^\dagger$,

$$\begin{aligned}
(\hat{u}^*(t), \hat{u}^*(t)) &\leqslant \frac{2}{\gamma}\frac{1}{2}(A^\dagger \hat{u}^*(t), A^\dagger \hat{u}^*(t)) \\
&\leqslant \frac{1}{\gamma}(A^\dagger \hat{u}^*(0), A^\dagger \hat{u}^*(0)) - \frac{1}{\gamma}\int_0^t (\hat{u}^*(s), B\hat{u}^*(s))ds.
\end{aligned} \tag{7.10}$$

Recall now the linear operator $T_B : L^2(U) \to L^2(U)$ and let e_k and λ_k be the eigenfunctions and eigenvalues, respectively, of the operator T_B.

We will next prove that $\lambda_k > 0$. Consider a function u which is non-zero everywhere. Then,

$$\begin{aligned}
(u, [T_B u]) &= \int_U \int_U u(x)B(x,y;\mu_0)u(y)dxdy \\
&= \int_\Theta \left(\int_U u(x)\sigma(wx+\eta)dx \right)^2 \mu_0(dcdwd\eta) + \int_\Theta \left(\int_U u(x)c\sigma'(wx+\eta)xdx \right)^2 \mu_0(dcdwd\eta) \\
&\quad + \int_\Theta \left(\int_U u(x)c\sigma'(wx+\eta)dx \right)^2 \mu_0(dcdwd\eta)
\end{aligned}$$

where $\mu_0(dcdwd\eta)$ assigns positive probability to every set with positive Lebesgue measure.

It is clear now that $(u, [T_B u]) \geqslant 0$. Suppose $(u, [T_B u]) = 0$. Notice that by [27], the fact that σ is bounded and non-constant by Assumption 2.3 implies that σ is also discriminatory in the sense of [13, 27]. Namely, if we have that

$$\int_U u(x)\sigma(wx+\eta)\pi(dx) = 0 \text{ for all } (w,\eta) \in \mathbb{R}^{d+1},$$

then $u(x) = 0$ in U. Then, $g(w,\eta) = \int_U u(x)\sigma(wx+\eta)dx = 0$ almost everywhere in $\mathbb{R}^{d+1}$ and by continuity and the fact that σ is also discriminatory we get that $g(w,\eta) = 0$ for all $(w,\eta) \in \mathbb{R}^{d+1}$. Then, $u(x) = 0$ everywhere in U by [13] . However, this is a contradiction, since u is non-zero everywhere. Therefore, if u is non-zero everywhere,

$$(u, [T_B u]) > 0. \tag{7.11}$$

Let us now consider the eigenfunction $e_k(x)$ with eigenvalue λ_k.

$$\begin{aligned}
(e_k, [T_B e_k]) &= (e_k, \lambda_k e_k) \\
&= \lambda_k (e_k, e_k) \\
&= \lambda_k.
\end{aligned}$$

Equation (7.11) implies that

$$(e_k, [T_B e_k]) > 0.$$

Therefore, we have indeed obtained that $\lambda_k > 0$. In addition, we can prove that $\lambda_k < C_\sigma < \infty$, i.e. the

eigenvalues are upper bounded.

$$
\begin{aligned}
(u, [T_B u]) &= \int_\Theta \left(\int_U u(x)\sigma(wx + \eta)dx \right)^2 \mu_0(dcdwd\eta) + \int_\Theta \left(\int_U u(x)c\sigma'(wx + \eta)xdx \right)^2 \mu_0(dcdwd\eta) \\
&\quad + \int_\Theta \left(\int_U u(x)c\sigma'(wx + \eta)dx \right)^2 \mu_0(dcdwd\eta) \\
&\leqslant \int_\Theta \|u\|_{L^2(U)}^2 \left(\int_U |\sigma(wx + \eta)|^2 dx \right) \mu_0(dcdwd\eta) + \int_\Theta \|u\|_{L^2(U)}^2 \left(\int_U |c\sigma'(wx + \eta)x|^2 dx \right) \mu_0(dcdwd\eta) \\
&\quad + \int_\Theta \|u\|_{L^2(U)}^2 \left(\int_U |c\sigma'(wx + \eta)|^2 dx \right) \mu_0(dcdwd\eta) \\
&\leqslant \int_\Theta \left(\int_U |\sigma(wx + \eta)|^2 dx + \int_U |c\sigma'(wx + \eta)x|^2 dx + \int_U |c\sigma'(wx + \eta)|^2 dx \right) \mu_0(dcdwd\eta) \times \|u\|_{L^2(U)}^2 \\
&= C_\sigma \|u\|_{L^2(U)}^2 ,
\end{aligned}
$$

Since $\sigma(\cdot)$ and U are bounded, we will have that $C_\sigma < \infty$. Under this assumption,

$$
\begin{aligned}
C_\sigma &\geqslant (e_k, [T_B e_k]) \\
&= \lambda_k.
\end{aligned}
$$

Let $\hat{u}(t, x) = \sum_{k=1}^\infty c_k(t)e_k(x)$ where e_k are the eigenvectors of the operator T_B (which form an orthonormal basis in L^2). The inequality (7.10) produces

$$
\sum_{k=1}^\infty c_k(t)^2 \leqslant C_0 - \frac{1}{\gamma} \sum_{i=1}^\infty \int_0^t \lambda_i c_i(s)^2 ds, \tag{7.12}
$$

where $C_0 = \frac{1}{\gamma}(A^\dagger \hat{u}^*(0), A^\dagger \hat{u}^*(0))$. Therefore,

$$
\begin{aligned}
c_k(t)^2 &\leqslant C_0 - \frac{1}{\gamma} \sum_{i=1}^\infty \int_0^t \lambda_i c_i(s)^2 ds \\
&\leqslant C_0 - \frac{\lambda_k}{\gamma} \int_0^t c_k(s)^2 ds.
\end{aligned}
$$

By Gronwall's inequality,

$$
\int_0^t c_k(s)^2 ds \leqslant C_0 e^{-\frac{\lambda_k}{\gamma} t}. \tag{7.13}
$$

Therefore, for any $\epsilon > 0$ and any $t > 0$, there exists a $\tau > t$ such that $|c_k(\tau)| < \epsilon$. We can use this fact to prove the following lemma.

Lemma 7.1. *For any function $f \in L^2$, any $\epsilon > 0$, and any $t > 0$, there exists a time $\tau > t$ such that*

$$
|(f, \hat{u}^*(\tau))| < \epsilon. \tag{7.14}
$$

Proof. Let e_k be the orthonormal basis corresponding to the eigenfunctions of the operator T_B. Since $f \in L^2$, it can be represented as

$$
f(x) = \sum_{k=1}^\infty b_k e_k(x).
$$

For any $\epsilon > 0$, there exists an M such that

$$
\sum_{k=M}^\infty b_k^2 < \frac{\epsilon^2}{2\|f\|^2}.
$$

27

Then,

$$(f, \hat{u}^*(t)) \quad = \quad (\sum_{k=1}^{M} b_k e_k, \hat{u}^*(t)) + (\sum_{k=M+1}^{\infty} b_k e_k, \hat{u}^*(t)).$$

Using the Cauchy-Schwarz inequality and the uniform bound (7.7),

$$\sup_{t \geqslant 0} \left| (\sum_{k=M+1}^{\infty} b_k e_k, \hat{u}^*(t)) \right| \quad \leqslant \quad \sup_{t \geqslant 0} \left(\sum_{k=M+1}^{\infty} b_k^2 \right)^{\frac{1}{2}} (\hat{u}^*(t), \hat{u}^*(t))^{\frac{1}{2}}$$

$$< \quad \frac{\epsilon}{2}.$$

For any t, there exists a $\tau > t$ such that $|c_k(\tau)| < \frac{\epsilon}{2M}$. Consequently,

$$(\sum_{k=1}^{M} b_k e_k, \hat{u}^*(\tau)) \left| \quad \leqslant \quad \sum_{k=1}^{M} |b_k c_k(\tau)| \right.$$

$$\leqslant \quad \sum_{k=1}^{M} (1 + b_k^2) |c_k(\tau)|$$

$$\leqslant \quad (1 + \|f\|^2) \sum_{k=1}^{M} |c_k(\tau)|$$

$$< \quad \frac{\epsilon}{2}. \tag{7.15}$$

Therefore, for any $t > 0$, there exists a $\tau > t$ such that

$$\left| (f, \hat{u}^*(\tau)) \right| < \epsilon. \tag{7.16}$$

$\square$

Recall the adjoint equation is

$$A^{\dagger} \hat{u}^* \quad = \quad (\langle u^* - h, m \rangle)^{\top} m. \tag{7.17}$$

Multiplying by the test function f from the set $\mathcal{F} = \{f \in L^2 : Af \in L^2\}$ and taking an inner product,

$$(f, A^{\dagger} \hat{u}^*) \quad = \quad (f, (\langle u^* - h, m \rangle)^{\top} m).$$

This leads to the equation

$$(Af, \hat{u}^*) \quad = \quad (f, (\langle u^* - h, m \rangle)^{\top} m).$$

Therefore, due to Lemma 7.1, we have the following proposition.

Proposition 7.2. *For any $\epsilon > 0$, any $f \in \mathcal{F}$, and any $t > 0$, there exists a time $\tau > t$ such that*

$$|(f, (\langle u^*(\tau) - h, m \rangle)^{\top} m)| < \epsilon. \tag{7.18}$$

Let $f = m_\ell$ in equation (7.18). Then, for any $\delta > 0$ and any $t > 0$, there exists a time $\tau > t$ such that

$$|\langle u^*(\tau) - h, m_\ell \rangle| < \delta. \tag{7.19}$$

Consequently, for any $\epsilon > 0$ and any $t > 0$, there exists a time $\tau > t$ such that

$$\langle u^*(\tau) - h, m \rangle^{\top} \langle u^*(\tau) - h, m \rangle < \epsilon. \tag{7.20}$$

28

8 Global Convergence of Pre-limit Objective Function

Define the pre-limit objective error function and limit error functions:

$$
\begin{aligned}
Q^N(t) &= (\langle u^N(t) - h, m\rangle^\top \langle u^N(t) - h, m\rangle)^{\frac{1}{2}}, \\
Q(t) &= (\langle u^*(t) - h, m\rangle^\top \langle u^*(t) - h, m\rangle)^{\frac{1}{2}}.
\end{aligned}
\tag{8.1}
$$

We can now prove the global convergence of the objective function $J^N(t)$.

Proof of Theorem 2.9. Due to equation (7.20), for any $\epsilon > 0$, there exists a time $T > 0$ such that

$$
Q(T) \leqslant \frac{\epsilon}{4}.
\tag{8.2}
$$

Due to Theorem 2.7, for any $\kappa > 0$, there exists an N_0 such that, for $N \geqslant N_0$,

$$
\mathbb{P}\left[\left\| u^*(T) - u^N(T) \right\| < \frac{\epsilon}{4(m^\top, m)^{\frac{1}{2}}} \right] > 1 - \kappa.
$$

We will select the following choice for κ:

$$
\kappa = \frac{\epsilon^2}{4C_J},
$$

where $\mathbb{E}\left[(Q^N(0))^2 \right]^{\frac{1}{2}} < C_J$. Let us take a moment to show this term is indeed uniformly bounded. Denoting with $(\cdot, \cdot)$ the usual inner product in L^2 and using the Cauchy-Schwarz inequality, Young's inequality, and a-priori elliptic PDE bounds based on the coercivity Assumption 2.1, we get that there is a constant $C < \infty$ that does not depend on N such that

$$
\begin{aligned}
\mathbb{E}\left[(Q^N(0))^2 \right] &= \mathbb{E}\left[\langle u^N(0) - h, m\rangle^\top \langle u^N(0) - h, m\rangle \right] \\
&\leqslant (m^\top, m)\mathbb{E}\left[\langle u^N(0) - h, u^N(0) - h\rangle \right] \\
&= (m^\top, m)\mathbb{E}\left[(u^N(0), u^N(0)) - 2(h, u^N(0)) + (h, h) \right] \\
&\leqslant (m^\top, m)\mathbb{E}\left[(u^N(0), u^N(0)) - 2(h, h)^{\frac{1}{2}}(u^N(0), u^N(0))^{\frac{1}{2}} + (h, h) \right] \\
&\leqslant (m^\top, m)\mathbb{E}\left[2(u^N(0), u^N(0)) + 2(h, h) \right] \\
&\leqslant (m^\top, m)\mathbb{E}\left[2C(g_0^N, g_0^N) + 2(h, h) \right].
\end{aligned}
\tag{8.3}
$$

By (4.7) in the proof of Lemma 4.3 we have that $\mathbb{E}\left[\left\| g_0^N \right\|^2 \right]$ is bounded. Therefore, we indeed have that

$$
\mathbb{E}\left[(Q^N(0))^2 \right] < C_J < \infty.
$$

By the triangle inequality and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
Q^N(T) &= (\langle u^N(T) - h, m\rangle^\top \langle u^N(T) - h, m\rangle)^{\frac{1}{2}} \\
&= (\langle u^N(T) - u^*(T) + u^*(T) - h, m\rangle^\top \langle u^N(T) - u^*(T) + u^*(T) - h, m\rangle)^{\frac{1}{2}} \\
&= \left\| \langle u^N(T) - u^*(T) + u^*(T) - h, m\rangle \right\|_2 \\
&= \left\| \langle u^N(T) - u^*(T), m\rangle + \langle u^*(T) - h, m\rangle \right\|_2 \\
&\leqslant \left\| \langle u^N(T) - u^*(T), m\rangle \right\|_2 + \left\| \langle u^*(T) - h, m\rangle \right\|_2 \\
&= (m^\top, m)^{\frac{1}{2}} \left\| u^N(T) - u^*(T) \right\| + \left\| \langle u^*(T) - h, m\rangle \right\| \\
&\leqslant (m^\top, m)^{\frac{1}{2}} \left\| u^N(T) - u^*(T) \right\| + \frac{\epsilon}{4}.
\end{aligned}
\tag{8.4}
$$

29

Consequently, for $N \geqslant N_0$,

$$\mathbb{P}\left[Q^N(T) < \frac{\epsilon}{2}\right] > 1 - \kappa.$$

Recall that $J^N(t) = \frac{1}{2}\left\langle u^N(t) - h, m\right\rangle^\top \left\langle u^N(t) - h, m\right\rangle$ is monotonically decreasing in t. Consequently, $Q^N(t) = \left(\left\langle u^N(t) - h, m\right\rangle^\top \left\langle u^N(t) - h, m\right\rangle\right)^{\frac{1}{2}}$ is also monotonically decreasing in t. Therefore, for $t \geqslant T$ and $N \geqslant N_0$,

$$
\begin{aligned}
\mathbb{P}\left[Q^N(t) < \frac{\epsilon}{2}\right] &= \mathbb{E}\left[\mathbf{1}_{Q^N(t) < \frac{\epsilon}{2}}\right] \\
&\geqslant \mathbb{E}\left[\mathbf{1}_{Q^N(T) < \frac{\epsilon}{2}}\right] \\
&= \mathbb{P}\left[Q^N(T) < \frac{\epsilon}{2}\right] \\
&> 1 - \kappa.
\end{aligned}
$$

Then, for $t \geqslant T$ and $N \geqslant N_0$,

$$
\begin{aligned}
\mathbb{E}[Q^N(t)] &\leqslant \mathbb{E}\left[Q^N(t)\mathbf{1}_{Q^N(t) < \frac{\epsilon}{2}} + Q^N(t)\mathbf{1}_{Q^N(t) \geqslant \frac{\epsilon}{2}}\right] \\
&\leqslant \frac{\epsilon}{2}\mathbb{E}\left[\mathbf{1}_{Q^N(t) < \frac{\epsilon}{2}}\right] + \mathbb{E}\left[Q^N(t)\mathbf{1}_{Q^N(t) \geqslant \frac{\epsilon}{2}}\right] \\
&\leqslant \frac{\epsilon}{2} + \mathbb{E}\left[Q^N(0)\mathbf{1}_{Q^N(t) \geqslant \frac{\epsilon}{2}}\right] \\
&\leqslant \frac{\epsilon}{2} + \mathbb{E}\left[(Q^N(0))^2\right]^{\frac{1}{2}}\mathbb{E}\left[\mathbf{1}_{Q^N(t) \geqslant \frac{\epsilon}{2}}\right]^{\frac{1}{2}} \\
&\leqslant \frac{\epsilon}{2} + \mathbb{E}\left[(Q^N(0))^2\right]^{\frac{1}{2}}\mathbb{P}\left[Q^N(t) \geqslant \frac{\epsilon}{2}\right]^{\frac{1}{2}} \\
&\leqslant \frac{\epsilon}{2} + \sqrt{C_J}\sqrt{\kappa}.
\end{aligned}
$$

Since we have selected $\kappa = \frac{\epsilon^2}{4C_J}$,

$$\sup_{t \geqslant T, N \geqslant N_0} \mathbb{E}[Q^N(t)] \leqslant \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

concluding the proof of the theorem.

$\square$

9 Extensions

In this work we studied global convergence for the optimization problem associated to the cost function given by (1.1). An interesting direction is to consider the same problem but for the cost function

$$J^N(\theta) = \frac{1}{2}\left\|u^N - h\right\|_{L^2(U)}^2, \tag{9.1}$$

where h is a targe profile. One can view the objective function (9.1) as a strong formulation of (1.1). In this case, the variable u satisfies the elliptic partial differential equation (PDE)

$$
\begin{aligned}
Au^N(x) &= f^N(x, \theta), && x \in U \\
u^N(x) &= 0, && x \in \partial U,
\end{aligned} \tag{9.2}
$$

and the adjoint PDE takes the form

$$A^\dagger \hat{u}^N(x) = u^N(x) - h(x), \qquad x \in U$$
$$\hat{u}^N(x) = 0, \qquad x \in \partial U. \tag{9.3}$$

The methods of this paper allow us to prove theorems analogous to Theorems 2.6 and 2.7. Consider the non-local PDE system

$$Au^*(t,x) = g_t(x), \quad x \in U$$
$$u^*(t,x) = 0, \quad x \in \partial U, \tag{9.4}$$

and

$$A^\dagger \hat{u}^*(t,x) = u^*(t,x) - h(x), \quad x \in U$$
$$\hat{u}^*(t,x) = 0, \quad x \in \partial U. \tag{9.5}$$

with

$$g_t(x) = -\int_0^t \left(\int_U \hat{u}^*(s,x')B(x,x';\mu_0)dx' \right) ds. \tag{9.6}$$

The system of equations given by (9.4)-(9.5) is well-posed as our next Theorem 9.1 shows.

Theorem 9.1. *Let as assume that Assumptions 2.1 and 2.2 hold. Then, there is a unique solution to the system (9.4)-(9.5) in $C([0,\infty); H_0^1(U)) \times C([0,\infty); H_0^1(U))$.*

Theorem 9.2. *Let $T > 0$ be given and assume the Assumptions 2.1, 2.2 and 2.3 hold. Then, as $N \to \infty$,*

$$\max_{0 \leqslant t \leqslant T} \mathbb{E} \left\| u^N(t) - u^*(t) \right\|_{H_0^1(U)} \rightarrow 0,$$
$$\max_{0 \leqslant t \leqslant T} \mathbb{E} \left\| \hat{u}^N(t) - \hat{u}^*(t) \right\|_{H_0^1(U)} \rightarrow 0,$$
$$\max_{0 \leqslant t \leqslant T} \mathbb{E} \left\| g_t^N - g_t \right\|_{L^2(U)} \rightarrow 0. \tag{9.7}$$

In addition, $\langle f, \mu_t^N \rangle \xrightarrow{p} \langle f, \mu_0 \rangle$ (convergence in probability) for each $t \geqslant 0$ for all $f \in C_b^2(\mathbb{R}^{2+d})$.

The proofs of Theorems 9.1 and 9.2 follow very closely those of Theorems 2.6 and 2.7 and thus will not be repeated here.

We can also prove the time-averaged weak convergence of $\hat{u}^*(t,x)$ and $u^*(t,x)$.

Theorem 9.3. *For any function $f \in L^2$ and any $\phi \in \{\phi' : A\phi' \in L^2\}$,*

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t (f, \hat{u}^*(s))^2 ds = 0,$$
$$\lim_{t \to \infty} \frac{1}{t} \int_0^t (\phi, u^*(s) - h)^2 ds = 0. \tag{9.8}$$

Proof. Similar to (9), it can be proven that

$$\int_0^t c_k(s)^2 ds \leqslant C_0 e^{-\frac{\lambda_k}{\gamma}t}. \tag{9.9}$$

Therefore,

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t c_k(s)^2 ds = 0. \tag{9.10}$$

Since $f \in L^2$, it can be represented as

$$f(x) = \sum_{k=1}^{\infty} b_k e_k(x).$$

For any $\epsilon > 0$, there exists an M such that

$$\sum_{k=M}^{\infty} b_k^2 < \frac{\epsilon^2}{2 \|f\|}.$$

Then,

$$(f, \hat{u}^*(t)) \;=\; (\sum_{k=1}^{M} b_k e_k, \hat{u}^*(t)) + (\sum_{k=M+1}^{\infty} b_k e_k, \hat{u}^*(t)).$$

Using the Cauchy-Schwarz inequality and the uniform bound (7.7),

$$\sup_{t \geq 0} \left| (\sum_{k=M+1}^{\infty} b_k e_k, \hat{u}^*(t)) \right| \;\leq\; \sup_{t \geq 0} \left(\sum_{k=M+1}^{\infty} b_k^2 \right)^{\frac{1}{2}} (\hat{u}^*(t), \hat{u}^*(t))^{\frac{1}{2}}$$

$$< \;\; \frac{\epsilon}{2}.$$

By the Cauchy-Schwarz inequality,

$$\left(\sum_{k=1}^{M} b_k e_k, \hat{u}^*(t) \right)^2 \;\leq\; \left(\sum_{k=1}^{M} |b_k c_k(t)| \right)^2$$

$$\leq \;\; \|f\|^2 \sum_{k=1}^{M} c_k(t)^2. \tag{9.11}$$

Then,

$$\frac{1}{t} \int_0^t (f, \hat{u}^*(s))^2 ds \;\leq\; \frac{C}{t} \int_0^t \left(\sum_{k=1}^{M} c_k(s)^2 + \epsilon^2 \right) ds. \tag{9.12}$$

Due to equation (9.10),

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t (f, \hat{u}^*(s))^2 ds \leq C\epsilon^2. \tag{9.13}$$

Since ϵ was arbitrary,

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t (f, \hat{u}^*(s))^2 ds = 0. \tag{9.14}$$

Consequently,

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t (\phi, u^*(s) - h)^2 ds \;=\; \lim_{t \to \infty} \frac{1}{t} \int_0^t (A\phi, \hat{u}^*(s))^2 ds$$

$$= \;\; 0. \tag{9.15}$$

$\square$

An open question for future research is to obtain a global convergence theorem for the pre-limit PDE $u^N(t, x)$ analogous to the global convergence result Theorem 2.9. This theorem would require proving a decay rate for $c_k(s)$, which remains a significant technical challenge. As we also mentioned in Section 1.1, another interesting future research topic is extending these theoretical results to nonlinear partial differential equations.

References

[1] H. Amann, Global existence for semilinear parabolic problems. J. Reine Angew. Math. 360 (1985), 47–83.

[2] T. Bosse, N.R. Gauger, A. Griewank, S. Günther, and V. Schulz, One-shot approaches to design optimization, Internat. Ser. Numer. Math. 165 (2014), pp. 43-66.

[3] C. Brandenburg, F. Lindemann, M. Ulbrich and S. Ulbrich, A continuous adjoint approach to shape optimization for Navier Stokes flow, in: Optimal Control of Coupled Systems of Partial Differential Equations, in: Internat. Ser. Numer. Math., vol. 158, Birkhäuser Verlag, Basel, (2009), pp. 35–56.

[4] A. Bueno-Orovio, C. Castro C, F. Palacios, and E. Zuazua. Continuous adjoint approach for the Spalart-Allmaras model in aerodynamic optimization. AIAA journal, 50(3), pgs. 631-646, 2012.

[5] J. Berg and K. Nystrom. Neural network augmented inverse problems for PDEs, arXiv preprint arXiv:1712.09685, 2017.

[6] M. Brenner, J. Eldredge, and J. Freund. A Perspective on Machine Learning for Advancing Fluid Mechanics. Physical Review Fluids, Vol. 4, 2019.

[7] S. Brunton, B. Noack, and P. Koumoutsakos. Machine Learning for Fluid Mechanics. Annual Review of Fluid Mechanics, Vol. 52, 2020.

[8] F. Cagnetti, D. Gomes, and H. Tran. Adjoint methods for obstacle problems and weakly coupled systems of PDE. ESAIM: Control, Optimisation and Calculus of Variations, 19(3), pgs. 754-779, 2013.

[9] A. J. Chorin, Numerical solution of the Navier–Stokes equations, Math. Comput. 22 (1968) 745–762.

[10] S. Cai, Z. Wang, F. Fuest, Y. Jeon, C. Gray, G. Karniadakis. Flow over an espresso cup: inferring 3-D velocity and pressure fields from tomographic background oriented Schlieren via physics-informed neural networks. Journal of Fluid Mechanics, 915, 2021.

[11] S Cai, Z Wang, S Wang, P Perdikaris, and G. Karniadakis. Physics-Informed Neural Networks for Heat Transfer Problems. Journal of Heat Transfer 43 (6), 2021.

[12] E. T. Curran, W. H. Heiser, and D. T. Pratt, Fluid phenomena in scramjet combustion systems, Annual Review of Fluid Mechanics 28, 323–360, 1996.

[13] G. Cybenko. Approximation by superposition of a sigmoidal function, *Mathematics of Control, Signals and Systems*, Vol. 2, pp. 303-314, 1989.

[14] K. Duraisamy. Perspectives on Machine Learning-augmented Reynolds-averaged and Large Eddy Simulation Models of Turbulence. arXiv:2009.10675, 2020.

[15] M. Duta, M. Giles, and M. Campobasso. The harmonic adjoint approach to unsteady turbomachinery design. International Journal for Numerical Methods in Fluids, 40(3-4), pgs. 323-332, 2002.

[16] L. Evans, *Partial differential equations*, American Mathematical Society, Volume 19, 1998.

[17] N. Gauger, A. Griewank, A. Hamdi, C. Kratzenstein, E. Özkaya, T. Slawig, Automated extension of fixed point PDE solvers for optimal design with bounded retardation, in: G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich, S. Ulbrich (Eds.), Constrained Optimization and Optimal Control for Partial Differential Equations, Springer, Basel, 2012, pp. 99–122.

[18] D. Gilbard and N.S. Trudinger, *Elliptic partial differential equations of second order*, 2nd edition, Springer, Berlin Heidelberg, 2001.

[19] M. Giles and N. Pierce. An Introduction to the Adjoint Approach to Design. Flow, Turbulence and Combustion, 65: 393–415, 2000.

[20] M. Giles and S. Ulbrich. Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 1: Linearized approximations and linearized output functionals. SIAM Journal on numerical analysis, 48(3), pgs. 882-904, 2010.

[21] M. Giles and S. Ulbrich. Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoint approximations and extensions. SIAM Journal on Numerical Analysis, 48(3), pgs. 905-921, 2010.

[22] S. Günther, N.R. Gauger and Q. Wang, Simultaneous single-step one-shot optimization with unsteady PDEs. Journal of Computational and Applied Mathematics, 294, (2016), pp. 12-22.

[23] F. H. Harlow and J. E. Welch, Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface, Physics of Fluids 8 (1965) 2182–2189.

[24] M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich, Optimization with PDE Constraints, in: Mathematical Modelling: Theory and Applications, vol. 23, Springer, New York, 2009.

[25] O. Desjardins, G. Blanquart, G. Balarac, and H. Pitsch, High order conservative finite difference scheme for variable density low Mach number turbulent flows, Journal of Computational Physics 227 (2008) 7125–7159.

[26] J. Holland, J. Baeder, and K. Duraisamy. Towards Integrated Field Inversion and Machine Learning With Embedded Neural Networks for RANS Modeling. AIAA SciTech Forum, 2019.

[27] K. Hornik. Approximation capabilities of multilayer feedforward networks. Neural Networks, 4(2), 251-257, 1991.

[28] S.B. Hazra, Direct Treatment of State Constraints in Aerodynamic Shape Optimization Using Simultaneous Pseudo-Time-Stepping, AIAA Journal, 45(8), (2007), pp. 1988-1997.

[29] S.B. Hazra, V. Schulz, Simultaneous pseudo-timestepping for PDE-model based optimization problems,Bit Numerical Mathematics, 44(3), (2004), 457-472.

[30] M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich, Optimization with PDE Constraints, in: Mathematical Modelling: Theory and Applications, vol. 23, Springer, New York, 2009.

[31] A. Jameson. Aerodynamic shape optimization using the adjoint method. Lectures at the Von Karman Institute, Brussels, 2003.

[32] A. Jameson, L. Martinelli, N. Pierce. Optimum aerodynamic design using the Navier-Stokes equations. Theoretical and computational fluid dynamics, 10(1), pgs. 213-237, 1998.

[33] A. Jameson and S. Kim. Reduction of the adjoint gradient formula in the continuous limit. In 41st Aerospace Sciences Meeting and Exhibit, pg. 40, 2003.

[34] X. Jin, S. Cai, H. Li, and G. Karniadakis. NSFnets (Navier-Stokes flow nets): Physics-informed neural networks for the incompressible Navier-Stokes equations. Journal of Computational Physics, 426, 2021.

[35] W. P. Jones and B. E. Launder. The prediction of laminarization with a two-equation model of turbulence, International Journal of Heat and Mass Transfer 15 (1972) 301–314.

[36] Y. Ju and W. Sun, Plasma assisted combustion: Dynamics and chemistry, Progress in Energy and Combustion Science 48 (2015) 21–83.

[37] L. Kaland, J.C. De Los Reyes, and N.R. Gauger, One-shot methods in function space for PDE-constrained optimal control problems, Optimization Methods & Software, 29(2) (2014), 376–405.

[38] E. Kharazmi, Z. Zhang, and G. Karniadakis. hp-VPINNs: Variational physics-informed neural networks with domain decomposition. Computer Methods in Applied Mechanics and Engineering, 374, 2021.

[39] J. Kim and P. Moin. Application of a fractional-step method to incompressible Navier–Stokes equations, Journal of Computational Physics 52 (1985) 308–323.

[40] J. Kim, P. Moin, and R. Moser. Turbulence statistics in fully developed channel flow at low Reynolds number, Journal of Fluid Mechanics 177 (1987) 133–166.

[41] D.A. Knopoff, D.R. Fernández, G.A. Torres and C.V. Turner. Adjoint method for a tumor growth PDE-contrained optimization problem, Computers and Mathematics with Applications, Vol. 66, (2013), pp. 1104–1119.

[42] D. Kochkov, J. Smith, A. Alieva, Q. Wang, M. Brenner, and S. Hoyer. Machine learning-accelerated computational fluid dynamics. Proceedings of the National Academy of Sciences, 118 (21), 2021.

[43] H. Kushner and G. Yin, Stochastic Approximation and Recursive Algorithms and Applications, Second Edition. *Springer*, 2003.

[44] O.A. Ladyženskaja, V.A. Solonnikov amd N.N. Ural'ceva, *Linear and quasilinear equations of parabolic type*, American Mathematical Society, Providence R.I., 1968, (Translations of Mathematical Monographs, Vol. 23, by S. Smith).

[45] B. E. Launder and D. B. Spalding, The numerical computation of turbulent flows, Computer Methods in Applied Mechanics and Engineering 3 (1974) 269–289.

[46] T.S. Lim, Y. Lu, and J. Nolen, Quantitative propagation of chaos in the bimolecular chemical reaction-diffusion model, SIAM J. Math. Anal., 52(2), (2020), pp. 2098-2133.

[47] J. Ling, R. Jones, and J. Templeton. Machine learning strategies for systems with invariance properties. Journal of Computational Physics, Vol. 318, 22-35, 2016.

[48] J. Ling, A. Kurzawski, and J. Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. Journal of Fluid Mechanics, Vol. 807, 155-166, 2016.

[49] J. F. MacArt, J. Sirignano, and J. .B. Freund. Embedded training of neural-network sub-grid-scale turbulence models. Physical Review Fluids 6 (2021) 050502.

[50] S. Nadarajah and A. Jameson. A comparison of the continuous and discrete adjoint approach to automatic aerodynamic optimization. In 38th Aerospace Sciences Meeting and Exhibit, pg. 667, 2000.

[51] S. Nadarajah and A. Jameson. Studies of the continuous and discrete adjoint approaches to viscous automatic aerodynamic shape optimization. In 15th AIAA Computational Fluid Dynamics Conference, pg. 2530, 2001.

[52] N. Pierce and M. Giles. Adjoint Recovery of Superconvergent Functionals from PDE Approximations. SIAM Review, 42(2), pgs. 247-264, 2000.

[53] P. P. Popov, D. A. Buchta, M. J. Anderson, L. Massa, J. Capecelatro, D. J. Bodony, and J. B. Freund. Machine learning-assisted early ignition prediction in a complex flow. Combustion and Flame, Vol. 206, 451-466, 2019.

[54] B. Protas. Adjoint-based optimization of PDE systems with alternative gradients. Journal of Computational Physics, 227(13), pgs. 6490-6510, 2008.

[55] M. Raissi, Z. Wang, M. S. Triantafyllou, and E. G. Karniadakis. Deep learning of vortex-induced vibrations. Journal of Fluid Mechanics, Vol. 861, 119-137, 2019.

[56] M. Raissi, P. Perdikaris, and E. G. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Fluid Mechanics, Vol. 378, 686-707, 2019.

[57] M. Raissi, H. Babaee, and P. Givi. Deep learning of PDF turbulence closure. Bulletin of the American Physical Society, 2018.

[58] M. Renardy and R.C. Rogers, *An Introduction to Partial Differential Equaitons*, Second edition, Texts in applied Mathematics, Springer-Verlag New York, Inc, 2004.

[59] J. Reuther, A. Jameson, J. Farmer, L. Martinelli, and D. Saunders. Aerodynamic shape optimization of complex aircraft configurations via an adjoint formulation. In 34th Aerospace Sciences Meeting and Exhibit, pg. 94, 1996.

[60] B. A. Sen and S. Menon. Linear eddy mixing based tabulation and artificial neural networks for large eddy simulations of turbulent flames. Combustion and Flame, Vol. 157, 62-74, 2010.

[61] B. A. Sen, E. R. Hawkes, and S. Menon. Large eddy simulation of extinction and reignition with artificial neural networks based chemical kinetics. Combustion and Flame, Vol. 157, 566-578, 2010.

[62] J. Sirignano, J. MacArt, and J. Freund. DPM: A deep learning PDE augmentation method with application to large-eddy simulation. Journal of Computational Physics, 2020.

[63] S. B. Pope, *Turbulent Flows*, Cambridge (2000).

[64] B. Steiner, Z. DeVito, S. Chintala, S. Gross, A. Paszke, F. Massa, A. Lerer, G. Chanan, Z. Lin, E. Yang, A. Desmaison, A. Tejani, A. Kopf, J. Bradbury, L. Antiga, M. Raison, N. Gimelshein, S. Chilamkurthy, T. Killeen, L. Fang, J. Bai, Pytorch: An imperative style, high-performance deep learning library, NeurIPS 2019.

[65] V. A. Sabelnikov and A. N. Lipatnikov, Recent Advances in Understanding of Thermal Expansion Effects in Premixed Turbulent Flames, Annual Review of Fluid Mechanics 49 (2017) 91–117.

[66] V. Srivastava and K. Duraisamy. Generalizable Physics-constrained Modeling using Learning and Inference assisted by Feature Space Engineering. arXiv:2103.16042, 2021.

[67] S. Ta'asan, One shot methods for optimal control of distributed parameter systems 1: Finite dimensional control. Technical report, DTIC Document, 1991.

[68] S. Ta'asan, Pseudo-time methods for constrained optimization problems governed by PDE, ICASE Report No. 95-32, 1995.

[69] M. Wang, A. Mani, and S. Gordeyev, Physics and Computation of Aero-Optics, Annual Review of Fluid Mechanics 44 (2012) 299–321.

[70] Z. Wang, K. Luo, D. Li, J. Tan, and J. Fan. Investigations of data-driven closure for subgrid-scale stress in large-eddy simulation. Physics of Fluids, Vol. 30, 2018.

[71] L Yang, X Meng, and G. Karniadakis. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. Journal of Computational Physics, 425, 2021.